

THE BALANCE OF SCALAR IMPLICATURE

by

Erin M. Tavano

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(LINGUISTICS)

December 2010

Copyright 2010

Erin M. Tavano

Dedication

For Elvis

Acknowledgements

My foremost and most affectionate thanks must go to Professor Elsi Kaiser and Professor Elena Guerzoni, who have been my mentors in all things throughout my graduate career. I would not have reached the end of it without their kindness and support.

I am also indebted to Professor Toben Mintz, Professor Roumyana Pancheva, and Professor Maria-Luisa Zubizaretta for their service on my committees and their valuable advice, comments and questions over the years, as well as Dr. Joyce Perez, who has so patiently guided me through the technicalities behind it all. Additionally, Professor Michael Arbib, Dr. Andrew Gordon, Dr. Jerry Hobbs, Dr. Ed Hovy, Dr. Jihie Kim, and Dr. Bonnie Glover-Stalls have together provided me with almost an entirely parallel education, which has been powerfully influential for me.

Comments from the participants of the Chicago Linguistic Society 46 and Penn Linguistics Colloquium 16, as well as from reviewers for the CUNY Conference on Human Sentence Processing 2008-2010, have contributed immensely towards this dissertation, as have members of the USC Psycholinguistics Lab Group.

Since 2002, my family has been repeatedly videotaped, transcribed, surveyed, eyetracked, and solicited for linguistic judgments. That any of them still talk to me at all is itself a source of wonder, and of course tremendous gratitude.

This dissertation owes its final existence to thirty-three kittens, five cats, and one beloved live-in statistics consultant, which is what the kids are calling it nowadays.

Table of Contents

Dedication	ii
Acknowledgements	iii
List of Tables	v
List of Figures	vi
Abstract	vii
Chapter 1: Introduction	1
Chapter 2: Eyetracking Experiment	18
Chapter 3: Lexical Decision Experiments	60
Chapter 4: Computational Model	99
Chapter 5: Conclusions and Future Research	122
Bibliography	135
Appendices:	
Appendix A: Chapter 2 Experiment Stimuli	140
Appendix B: Chapter 3 Experiment Stimuli	142

List of Tables

Table 1: Chapter 2 Experiment Stimuli	32
Table 2: Chapter 2 Participant Mean Response Times by Main Factors	43
Table 3: Chapter 2 Participant Mean Response Times by Condition	43
Table 4: Word Association Proportions for Common GCI Scales	65
Table 5: Average Number of Words Linked to Prime Words (SIAM)	114
Table 6: Ranking of E/A and NE/A Target Word Nodes (SIAM)	116
Table 7: Number of Nodes from which Target Words Are Linked (SIAM)	117
Table 8: Chapter 2 Experiment Target Items	140
Table 9: Chapter 2 Experiment Filler Items	140
Table 10: Target Words Used in Experiments 1-3 in Chapter 3	142
Table 11: Chapter 3 Experiment Nonwords	142
Table 12: Basic Contexts Used in Experiments 1 and 2 in Chapter 3	143
Table 13: Implicature Contexts Used in Experiments 1 and 2 in Chapter 3	144
Table 14: Additional Text for Lengthened Contexts in Chapter 3 Experiment 3	145
Table 15: Filler Items for Chapter 3 Experiments 1-3, and Associated Target Items	147

List of Figures

Figure 1: Storto and Tanenhaus (2005) Sample Scenes for "And"	26
Figure 2: Storto and Tanenhaus (2005) Sample Scenes for "Or"	26
Figure 3: Huang and Snedeker (2009) Sample Visual Scene	27
Figure 4: Sample Stimuli for Chapter 2 QAPA and QSPS Conditions	34
Figure 5: Sample Filler Images and Sentences	35
Figure 6: Participant Responses	40
Figure 7: Participant Responses	43
Figure 8: Proportion of Inspections on Set of Target / Nontarget Items	47
Figure 9: Chapter 3 Experiment 1 Response Times in Each Target Word Condition, First Trials	77
Figure 10: Chapter 3 Experiment 1 Response Times in Each Target Word Condition, First and Second Trials	78
Figure 11: Chapter 3 Experiment 2 Response Times in Each Target Word Condition, First Trials	87
Figure 12: Chapter 3 Experiment 3 Response Times in Each Target Word Condition, First Trials	93
Figure 13: How SIAM Spreads Activation to New Nodes	111
Figure 14: Sample Visual Scenes from Huang and Snedeker (2009) and Grodner et al. (2010)	127

Abstract

This dissertation aims to provide new insight into the properties that underlie scalar implicature (SI) processing. Previous experiments have investigated the time course of SI processing and whether or not it is a costly, resource-demanding process, but not what the resources have been specifically used to do, and when. Additionally, word scales used in SI research have been treated as interchangeable variations on a type, without systematic evaluation of that assumption. This dissertation investigates these questions via a computational model and several experiments using highly sensitive methodologies, including eyetracking. It concludes that scalar implicature is a costly, effortful process, though that effort is easily affected by task or other experimental factors. Specifically, processing difficulty appears to arise locally (as the scalar term is encountered) and the effort put towards a scalar inference, where the higher term(s) on the scale are negated. This negation appears to lead to reduced accessibility of the term.

Additionally, it is not clear that the scales must be represented as such in the lexicon, or that they should be treated uniformly regardless of the words that compose them. Rather, this research shows that "typical" scales, the ones most often discussed in SI research, are composed of words that are both strongly and unidirectionally connected in the mental lexicon, in the order of the scale. In comparing many scales, I conclude that entailment is neither necessary nor sufficient in a scale, and scalar implicature may involve any set of related and sufficiently frequent words. This points toward a future unification account of particularized and generalized scalar implicature.

Finally, my findings also have methodological implications for experimental work in scalar implicature. The results suggest that participants' responses in typical experimental tasks, such as judging whether a sentence is true or appropriate, can be delayed not only by implicature processing but also (for example) participants' evaluation what the experimenter considers to be appropriate.

Chapter 1: Introduction

Psycholinguistic investigation is challenging when we study what is said, and even more so when we consider what is not said. In recent years, researchers have been embracing this challenge in increasing numbers. Experimental work on the processing of conversational implicature has begun to be undertaken in earnest. In this dissertation, I continue this research by examining some questions and assumptions of implicature processing, which have not yet been investigated in detail, but which are important to both psycholinguistics research and pragmatics theory.

Implicature

Grice (1975) introduced the idea that people seek to be maximally helpful, informative, and relevant when conversing with others. This idea constitutes his Cooperative Principle: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged" (45). The behaviors required by the principle are further detailed in his Conversational Maxims. To summarize the first three briefly:

1. Maxim of Relevance (Say only what is relevant)
2. Maxim of Quality (Say only what is true)
3. Maxim of Manner (Be clear)

Finally, there is the maxim of primary importance to this work, the Maxim of Quantity, which will be addressed in detail later.

"1. Make your contribution as informative as is required (for the current purposes of the exchange.)

2. Do not make your contribution more informative than is required."

(Grice 1975:45)

According to Grice, speakers generally make the effort to obey the conversational maxims, and hearers generally assume that speakers are doing so. Mutual awareness of the Cooperative Principle is key to expressing a meaning without saying it. For example, someone may ask me how my dissertation is going, and I would respond, "Well, I'm still alive." By this indirect answer, I intend to convey that the progress is difficult but going as well as could be expected. Because I believe that the questioner believes that I am providing a reasonable answer to the question, I expect him to infer its intended meaning. In general, when a speaker intentionally says something irrelevant, unclear, blatantly false, or notably under- or over-informative, with the purpose of conveying a nonliteral meaning, she is said to have created an implicature. The hearer, believing in the speaker's default commitment to cooperation, infers that such an implicature exists when he hears what he judges to be an "uncooperative" utterance. It is then up to him to infer the intended meaning.

The processing leading up to that inference is the present concern. For instance, much previous research has addressed the question of whether or not understanding

implicature is a costly, resource-demanding process, that is, whether it is effortful for the hearer to infer that the speaker has intended to express some unspoken meaning, and what that meaning might be. Another open question regards the time course of processing. How much of an utterance must be processed, and to what degree, before pragmatic inferences are made? As these two issues have been most often addressed, it is necessary to discuss their specific motivation.

There are a variety of reasons implicature processing is thought to be difficult. For instance, in the case of what Grice calls a Particularized Conversational Implicature (PCI), considerable integration of contextual or world knowledge is necessary for the hearer to derive the speaker's unspoken meaning. In contrast, the Generalized Conversational Implicature (GCI) is a type of meaning that any speaker typically intends to convey when using a particular word, even though it is not part of that word's meaning. The following exchange illustrates both implicature types.

- (1) A: "Where's John?"
B: "Some of the guests are already leaving."
PCI: "Perhaps John has already left".
GCI: "Not all of the guests are already leaving". (Levinson 2000:17)

Because of the Cooperative Principle, A will assume that B intends him to understand that B's answer will provide A with the information he needs. If A has the particular contextual knowledge that John is a guest and that there is a party going on, as well as the general world knowledge that guests do not stay forever at parties, he infers

the above PCI. Even if A does not have this information he might infer it solely from the apparent irrelevance of the response, though perhaps with less certainty. However, A does not need to know anything about John, guests or parties to infer the GCI. It is commonly assumed that the ability to make this often-repeated inference comes with competence in the language.

Some time and effort is generally imputed to the above (and any) PCI, though it may vary with the length of the required chain of inference. The relative difficulty of integrating many different pieces of information, in order to compute a conclusion of variable certainty, is uncontroversial. The question of the current study, then, is what type of processing is called for by the GCI, seemingly standardized in language, very frequent, perhaps even automatic, but also seeming to require some kind of inference.

Scalar implicature

The source of the above "Not all" implicature is the entailment scale, or Horn scale (Horn 1972) on which we as language speakers understand "some" to be placed, typically $\langle \text{all}, \text{some} \rangle$, where the leftmost item ("all") is the highest on the scale, descending rightward, such that each higher item entails the one below it. A specific interpretation of the Maxim of Quantity requires the speaker to choose the appropriate term on such a scale, to the best of her knowledge, to be as informative as possible but no more informative than she can be. This is known as *scalar implicature*. If the hearer believes the speaker is obeying Quantity, he is expected to infer that by choosing "some" and not "all", the speaker intends to convey that the stronger or higher item on the scale is either not true,

or the speaker does not know or have evidence that it is true. Therefore, if she has said, "Some of the guests are already leaving", she intends the hearer to infer "Not all of the guests are already leaving."

The idea of scales is complicated by the fact that many sets of words, which are not related by logical entailment, nonetheless serve as scales for scalar implicature. A common example is <succeed, try>. "Succeed" does not entail "try" – a person can succeed without trying – and yet it is very common for a speaker to say that she "tried" to do something, intending to convey the meaning that she did not succeed. In spite of the lack of an entailment relation, this is still thought of as scalar implicature, in part because it meets Grice's cancellability requirement for implicature: If the speaker wanted to express that she had succeeded, she would have had to explicitly deny the implication that she had not succeeded. The necessity of doing so is evident in the examples below.

- (2) a. "I tried to make a cake."
- b. "I tried to make a cake, and in fact I succeeded."

Although the scales in scalar implicature are usually thought of as lexical (and are defined as such by Horn (1972), Gazdar (1979) and Levinson (2000)), units greater than a word have been discussed in the same terms. For instance, conjunctive phrases have an entailment relationship. In general, some property that holds for X and Y together also holds for X and Y individually. Given some conjunct, a speaker that makes a claim about one part of it may be negating that claim for the other part.

- (3) A. "Do you know if this bus goes to Washington and Lake?"
B. "It goes to Washington...."

The clear meaning is that B does not know if the bus goes to Lake (or that B knows that it does not, though if B were able to make this stronger statement, he would normally have done so.) This implicature must be said to derive from the Maxim of Quantity, even if it is not exactly scalar, as it exploits the idea that a speaker should say no more than she can say. Furthermore, the implicature is generalized. No specific knowledge about buses or neighborhoods is necessary to understand its meaning.

Even further afield are implicatures that are intuitively feel scalar, but are neither lexical nor general. Such scales can be generated by context. For example:

- (4) Q. Do you speak Greek?
A. I like Greek food... (Hirschberg 1985)

The hearer will infer from this that the respondent does not speak Greek, having given the most information possible about his or her Greek-related competence. The speaker and hearer are apparently aware of an ad-hoc ordered scale <speaking Greek, liking Greek food>. It is clear that these predicates are ordered, as there is something "less" about liking Greek food relative to speaking Greek. The ordering principle is not entailment, but something less well defined, possibly degree of appreciation of Greek culture. Liking Greek food seems reasonable as a landmark on the way to learning to speak Greek, which itself might be a stop on the way to Greek citizenship, and so on. Of

course this scale of Greek appreciation is not something all speakers of a language are aware of. Rather, it has been created on the fly from world knowledge, and the recently mentioned predicates.

The present work, along with most of the related previous psycholinguistic research, addresses scalar implicature specifically and not implicature from the other maxims, and addresses lexical entailment (Horn) scales and not other types. This is because this domain of scalar implicature has the apparent benefit of clear predictions, uncomplicated by context or world knowledge. As we have seen, though, scalar implicature is so easily extensible beyond entailment scales that it seems quite inadvisable to ignore the other varieties, as difficult as it can be to reach broadly applicable conclusions. While it may be a practical necessity for experimentation and modeling to limit the data to lexical items, the work in this dissertation aims for results that are generalizable across scales of different types, and applicable to Quantity implicature of predicates and propositions.

Open questions

We have just introduced the first of the many open questions about scalar implicature, namely, what may constitute a scale? This question is usually treated theoretically, and the work herein is, to my knowledge, the first experimental approach to it. However, given the above understanding of implicature, psycholinguistics researchers have sought to explain the processing behind it, and I attempt to contribute to these issues as well.

Experimental research has addressed two main processing issues behind scalar implicature. The first is, to what degree is it effortful for the hearer. Inferences drawn from a speaker's scalar implicature are very fast, seemingly automatic. However, whether the inference is *actually* drawn or not is highly dependent on context, and perhaps other factors such as cognitive load. The second question relates to the time course of implicature processing. Exactly when does a hearer draw the inference from implicature? Is it immediately on hearing a "lower" scalar word, or must the semantics of the entire proposition be resolved first? The following sections address the theoretical accounts that underlie these issues.

Theoretical background

The Default view, and Relevance Theory

The Gricean maxims have been reformulated in various ways under what is known as the neo-Gricean program, in efforts to (among other things) resolve overlap between maxims, extend the account towards other linguistic phenomena such as pronouns, and to explicate the mental processes that underlie implicature. The neo-Gricean account of implicature that concerns us here is often called the Default view, and is mainly from Levinson (2000). The account can be outlined as follows: Imagine if the respondent (B) in example (1) had to say, "John, who is a guest at the party that is currently going on, may have already departed the party, like at least some of the other guests are doing though not all of them." It would be unnecessarily time-consuming to use speech to

transmit all the information a speaker intends to convey. Thus people have developed or evolved heuristics, like implicature, to quickly and fully understand what is communicated without the need for the speaker to spell it out.

As we have briefly noted above, implicatures are not always processed for one reason or another. Under the Default view, when an expected inference does not seem to arise, it is because it has been cancelled by its inconsistency with the common ground of the discourse.¹ In other words, by default, the inference is always drawn, and then sometimes cast aside. Although Levinson does not make very specific processing claims, it is understood that, under this view, the processing of implicatures is fast and essentially effortless. This holds only for generalized implicatures. Particularized (context-dependent) implicatures, while having many apparent similarities, are of a fundamentally different nature, and it is wrongheaded to search for a single basis for both.

The account in primary opposition to the Default view is Relevance Theory, introduced by Sperber and Wilson (1995), which treats implicature processing as any other inference, giving no particular favor to its linguistic status. Relevance Theory states that people seek to be efficient in all their cognition, linguistic and otherwise. To that end, people will generally only process communication that is the most relevant, that is, promises to yield the most new information when combined with a context, for the least processing cost. A presupposition of this is, of course, that implicature processing does bear a cost, which may at times be more than the hearer can cognitively afford. Furthermore, Relevance Theory implies that there is no reason to distinguish the

¹ The processing status of inference cancellation has not been much discussed; Levinson seems to indicate that the inference merely never happens, though some psycholinguists (e.g. Bezuidenhout and Morris 2004) have assumed that a reading slowdown is indicative of effortful cancellation.

processing mechanisms for generalized and particularized implicatures, except that the latter may be more costly than the former. It also explicates the role of context in implicature.

Perhaps the least intuitive part of Relevance Theory is the new role of context as something that is not given, but rather determined or chosen by the hearer. In other words, while other theories assume that an utterance is processed against information from an existing context, Relevance Theory claims that the context in which to understand an utterance is produced by the hearer in response to that utterance. This begins to sound like a chicken-and-egg problem: how can a hearer understand an utterance well enough to choose a context for it, without already knowing its context? However, “relevance” is a concept intended to be a deterministic link (for a given individual at a given time) from utterance to context. An utterance is relevant to an individual based on the degree of two input factors: 1) the magnitude of “contextual effects” of an utterance (briefly, the degree to which the hearer's knowledge store would be altered), and 2) the degree of effort required by the hearer. The hearer seeks to maximize the first and minimize the second. In the attempt to balance these interacting factors, the hearer decides whether it is appropriate to draw an inference in a particular context.

Relevance Theory thus reduces the Gricean program to the one maxim for which the theory is named. Even the overarching Cooperative Principle is unnecessary: a sense of cooperation, if present, is merely part of the context. Again, in presuming that inference requires effort, sometimes prohibitive, even for GCIs, the Relevance Theory view opposes the Default view.

Sperber and Wilson note that there are many details of Relevance Theory yet to be addressed. Levinson (1989; 2000), on the other hand, seems to believe that failures of the account are quite damning. In fact, there are many points which may be addressed in comparing these two accounts, but to summarize the ones pertinent to the current study, we may say that Relevance Theory assumes that every inference is costly and only made if the hearer thinks it is worth the cost, and the Default view assumes that GCIs, at least, are automatically inferred and not costly. Sperber and Wilson argue that the Default view is insufficiently generalized and not psychologically realistic. Levinson, on the other hand, argues that Relevance theory lacks predictive power (as it depends on so many factors, including the individual), and unable to explain why language processing is so fast if so many inferences must be generated every time even when they are apparently standardized. The language-specific nature of the Default view need not be a disadvantage. For instance, in Levinson's formulation of the Gricean maxims, patterns of lexicalization for scalar items are predicted in a way that they would not be on a general psychological account. It stands to reason that given the imbalance in speed between cognition and linguistic expression, language processing in the brain might have evolved in very specific ways that do accord with neo-Gricean maxims

Time course of processing

Another point of contention involves the timing of implicature processing, or pragmatic processing in general, relative to semantics and syntax. The two main views on this issue are known as the Global and Local accounts, and mainly concern whether it is serialized between independent modules, each charged with a processing function (e.g., syntax, followed by semantics), or whether the functions work in parallel or are otherwise integrated in a non-linear way.

In the Global account, implicatures are processed subsequent to semantic parsing. A grammatical module translates the input utterance into logical form, and passes it to a pragmatics module, which applies factors to yield scalar implicatures. That is, implicature processing must apply to entire utterances, and only after its grammatical meaning has been computed. Levinson's Default view assumes this approach, as do Geurts (2007) and Russell (2006). Such modularity must sound like very old-style thinking to many psycholinguists and probably to theorists as well, as it has been shown many times over that pragmatic information can guide parsing or override what would seem to be standard parsing procedure (e.g., Altmann & Steedman 1988, Steedman & Altmann 1989; also summary in Pytkänen and McElree 2006).

The Local account of implicature generation, as in Chierchia (2004, 2006) and Fox (2007) is intended to resolve problems arising from this idea of sequential interpretation. Two often-cited ones are incorrect predictions for scalar items in embedded clauses (“John believes that some students are waiting for him” → “It is not the case that John believes that every student is waiting for him”) (Chierchia 2004:44), and

scalar items under disjunction (“Mary is either working at her paper or seeing some of her students” → “It is not the case that Mary is either working at her paper or seeing all of her students” → “Mary is not working at her paper”) (Chierchia 2004: 46) The debate involves many other phenomena, but these two issues are sufficient to motivate the Localist view for now.

These two Localist accounts both work as follows. There is some type of operator, similar in denotation to the word “only”, inserted at points within the logical form. The operator generates alternatives to the scalar item as the LF is being computed. In other words, rather than completely parse the utterance and then pass it to a pragmatic module, as in the global account, the grammatical module partially parses the utterance, receives input from the pragmatic module via the operator, then returns it to the grammatical module along with the implicatures (that is, the negated stronger version(s) of the proposition just parsed). The grammatical module then goes on creating the LF of the utterance now using the strengthened meaning, repeating the process as necessary. Among the objections to the Local accounts are that they confuse semantics and pragmatics, and that they still have empirical problems, but this brief overview is all that is necessary for the present.

Psycholinguists have not addressed many specifics of either the Global or Local accounts of implicature processing, as it is quite hard to design experiments that measurably distinguish between the accounts while excluding other sources of any effects. Research so far has assessed whether implicature processing takes additional time or not, but not what goes on in that time. While the present work has the same

constraints as previous research, it does contribute some evidence in favor of the localist view (particularly Chapter 2).

Overview

This goal of this dissertation is to add new insight into the properties that underlie scalar implicature processing. For instance, previous experiments have addressed the question of whether or not understanding implicature is a costly, resource-demanding process, but have not addressed what those resources are specifically required to do. The time course of implicature has similarly been investigated, without in-depth research into the nature of processing activity during the given time. Finally, word scales have been treated as interchangeable variations on a type, without systematic evaluation of that assumption, or the degree to which hearers are aware of the particular scales most often used in experiments. This dissertation investigates all of these questions, concluding that:

1. Scalar implicature is clearly a costly, effortful process within the contexts investigated herein, though that effort is easily confounded by experimental factors.
2. Processing difficulty appears to arise locally (at the scalar term) and is put towards a scalar inference, where the higher term(s) on the scale are negated.

3. There is additional evidence that whether or not an implicature is processed depends on the hearer's cognitive load, and the supportiveness of the context; so it is not automatic, though could be considered "default".

4. It is not clear that the scales must be represented as such in the lexicon, or that they should be treated uniformly regardless of the words that compose them. Rather, the evidence suggests that there is inhibition of any related words in implicature-supportive contexts, not necessarily scale words. This points toward a future unification account of particularized and generalized scalar implicature.

In addition, I discuss important considerations of methodology in scalar implicature experimentation, which have not been well-understood in previous research.

Participants' responses in typical experimental tasks, such as judging whether a sentence is true or appropriate, can be delayed not only by implicature processing but also (for example) participants' evaluation what the experimenter considers to be appropriate. For these and other reasons (e.g., possible priming of scalar inferences, learning effects), it is common to find that the magnitude of effects varies over repeated trials in a condition, and experimenters must be careful to evaluate individual trials within a condition.

Chapter summary

Following this introductory chapter, this dissertation is structured as follows.

Chapter 2 presents the results of an eyetracking experiment which uses a new methodological variation to illustrate underlying processing during the interpretation of a scalar implicature. This experiment finds that implicature processing does incur a cost, which is quantifiable in eyetracking measures as well as time. Additionally, the results show that the measurable effects of implicature processing can depend not only on context, but also the experiment task itself. The questions asked by the experimenter and the repetition of stimuli that require implicature processing may both significantly affect results.

Chapter 3 reports on a series of experiments that investigate the nature of scales. First, my research shows that "typical" scales, the ones most often discussed in scalar implicature, are composed of words that are both strongly and unidirectionally connected in the mental lexicon, in the order of the scale. For instance, with regard to a scale like <hot, warm>, where "hot" entails "warm", participants in a free-association task will very often produce the word "hot" as the first word they think of when they hear "warm", but not vice-versa. This suggests that an associative connection may be additionally contributing to scalar implicature, as well as the entailment relationship. The experiment compares many scales, with components manipulated by entailment and associate status. I conclude that entailment is not a necessary or sufficient property for scales. Rather, processing effects from scalar implicature are most apparent for high-frequency, entailing, associated word scales, but also for other scale types.

Chapter 4 presents the Scalar Implicature Activation Model (SIAM), a computational connectionist model of how implicature processing might be implemented in the lexicon. SIAM models a lexicon of both semantic and more generally associative

links to nodes which represent both word forms (that is, strings of letters) and concepts. Using the contextual stimuli from the Chapter 3 experiments as input, SIAM demonstrates that something like scalar implicature can arise from an organization of the lexicon that does not specifically implement scales. If a mental structure for scales, and scalar implicature, is not needed, this suggests that the effects of scalar implicature are, as Relevance Theory suggests, generalizable to inference over knowledge as a whole. This provides motivation to restart efforts into unifying generalized and particularized scalar implicature.

Chapter 5 concludes this dissertation with a summary of results, some implications for theoretical results, and suggestions for future research.

Chapter 2: Eyetracking Experiment

This chapter reports the results of an eyetracking experiment that explores the underlying processing of scalar inferences invoked by Generalized Conversational Implicatures. The study is designed to explore the differences between the Default view of automatic, costless scalar implicature processing (Levinson 2000), and the Relevance Theory view (Sperber and Wilson 1995), which suggests that implicature processing involves a costly inferential process. In this experiment we seek evidence for one or the other of these two positions, with regard to the scalar item "some". On hearing "some", is "not all" automatically, instantly and costlessly calculated, perhaps cancelled later, or only computed in a context that is relevant, like inferences in general?

Although the experiments in this chapter do not fully support either of these approaches, they suggest that implicature processing does entail a cost, at in the kind of situation presented here, where linguistic stimuli is paired with a visual context. This cost is evidenced by longer response times, and also by eye movements, which increase in frequency after hearers process implicatures. An increase in eye movements is also present immediately after hearing a word that is lower on a scale (e.g., "some" in the scale <all,some>). This provides support for Localist theories of implicature. Finally, I discuss data analysis and methodological concerns that are applicable to future experiments in implicature.

Previous experimental work

Many techniques have been utilized in the study of scalar implicature processing, and all have controversial aspects. Most experiments offer response times (of a judgment, or in self-paced reading) as an indicator of processing load, and generally any observed effort on processing a scalar inference is taken as supporting the Relevance view (i.e., that some effort is required). ERP and eyetracking studies offer the hope of more insight into online processing, and sometimes a more naturalistic experimental environment, but there is debate about the specific implementation of these methods and the meaning of their results. As the current study uses eyetracking, we will first look at other online experiments using different methodology, then turn to a more detailed look at other eyetracking studies on implicature.

Non-eyetracking studies

Noveck and Posada (2003) conducted an ERP study on implicature processing, evaluating both response time and the N400 brainwave response across conditions. The N400 wave indicates semantically anomalous stimuli; a larger N400 response indicates greater perceived anomaly. Participants were asked to give truth judgments on three types of sentences:

(5)

- a. *True sentences* "Some houses have bricks"
(But not all of them do)

- b. *False sentences* "Some birds have televisions"
(None of them do, at least in a real world context.)

- c. *Underinformative sentences* "Some elephants have trunks"
(The judgment depends on a scalar inference. All elephants have trunks, so the sentence is not giving as much information as it could give.)

Like a number of other studies on scalar implicature processing, Noveck and Posada demonstrated a split among their participants with regard to the interpretation of the weaker item on the scale (here, "some"). Participants tend to give either a consistent pragmatic interpretation, that is, influenced by the scalar implicature (answering "False" to "Some elephants have trunks", since all elephants have trunks), or a consistent logical interpretation of "some" (answering "True", since some elephants *do* have trunks). The pragmatic interpretation group was slower than the logical interpretation group, and the delay was attributed to the time taken to (perhaps electively) process the scalar implicature. This is debatable and more will be said about the nature of participant response groups later.

With regard to the ERP results, Noveck and Posada looked at the N400 response to the last word of each sentence, and found that underinformative items yielded a lower response relative to the obviously true and false items, indicating relatively low perceived semantic anomaly. Further, there was no difference between the participant interpretation groups. Combined with the longer response time, the authors attribute this to a pragmatic process (or other decision-making process) that took place after the last word, thus better supporting the Relevance Theory view.

Bott and Noveck (2004) used stimuli similar to Noveck and Posada (2003) to collect truth judgment response times, while controlling for some potential concerns in the earlier experiment. Additional variations included modified instructions (mentioning explicitly that "some" could mean "some and not all"), varying whether participants gave positive or negative responses to underinformative sentences such as (5c), and fixing the allowed response time. The overall results were similarly suggestive of cost and, therefore, Relevance. When longer response times were allowed, there were more pragmatic interpretations of underinformative sentences, that is, participants were more likely to reject the truth of a sentence like "Some elephants have trunks."

Also supporting the Relevance view are Breheny, Katsos, and Williams (2006) and Katsos, Breheny and Williams (2005), who conducted self-paced reading studies (in Greek), showing that the time to read a segment containing "some" was longer when the implicature was relevant to the context, compared to when it was not. They also created neutral contexts, i.e. without indicating whether the implicature was relevant, by varying the sentence position of a quantifier phrase (as in example 6 below).

- (6) a. "Some of the consultants had a meeting with the director.

The rest did not manage to attend."

→ Effortless if implicature is accommodated

- b. "The director had a meeting with some of the consultants.

The rest did not manage to attend."

→ Effortful if implicature is processed late (under Relevance), or
effortless (under Default) (Breheny, Katsos, and Williams 2006:447)

If the phrase appeared sentence-initially (6a), the experimenters believed that comprehenders would treat it as old information (a tendency reportedly strong in Greek) and accommodate the "not all" inference as part of the context. Then, when they heard a second sentence referring to "the rest" of the consultants, there should be no reading slowdown as the idea of there being remaining consultants is already present. However, if the quantifier phrase appeared late in the sentence (6b), no contextual accommodation would be made, and it should be effortful to make an inference about the remaining consultants. Under the Default view, however, the position of the quantifier phrase should be irrelevant, as it would automatically trigger an inference in either case.

The results showed that the position manipulation was effective in causing or preventing an inference. Thus, the authors conclude that scalar inference is a local process (i.e. calculated as the quantifier is heard, without needing to wait for the end of the sentence) but not default, as it only arises in supportive contexts.

Although the previous experiments have supported the idea of costly inference, Foppolo (2007) presents a different perspective. Her experiment, a picture verification task with response-time measures, uses another type of scale, that of the logical connectives <and, or>. In natural language (as opposed to logic), "or" usually has an exclusive interpretation, that is, saying "A or B" means "A or B but not A and B". The *exclusive or* is thought to be derivable from the scalar implicature inherent in saying "or" rather than the stronger "and". However, that implicature may sometimes be cancelled, yielding the *inclusive or*, meaning "A or B and no information about whether A and B".

The experiment in Foppolo (2007) was designed to test a middle-ground hypothesis from Chierchia (2006), where scalar inferences are indeed costly to calculate, but only in certain structural contexts. Her experiment specifically examines Downward Entailing (DE) contexts, which are represented in the experiment by the antecedent clause in a conditional statement: If X **or** Y, then Z. The comparably easy, presumably costless context is where the scalar item is in the consequent: If Z, then X **or** Y. Participants read one of these sentences then looked at a scene divided into four mini-scenes, which, taken together, represented the logical (inclusive "or") interpretation or pragmatic (exclusive "or") interpretation. Participants had much lower acceptance rates for the pragmatic (exclusive "or") interpretation in the DE context, and were significantly slower to accept it than reject it. No other response time comparisons were significant. The response time results stand in interesting contrast to other studies that have found slower response times for scalar inferences, but Foppolo suggests that it may not be the

inference generation that has cost, but rather that the difficulty may be due to the reduced informativity when the implicature is included. ²

Eyetracking studies

Eye movements have been used to study language as early as Cooper (1974), but only recently have technological advances in eyetrackers facilitated their widespread use. This development has been utilized in both silent-reading experiments, and experiments in the visual-world paradigm, where participants look at a visual representation of a scene as they hear spoken language (as in the present experiment). Eyetracking has been a real benefit for psycholinguistic studies, as eye movements are highly sensitive, low-cost, low-threshold indicators of attention, or concept activation (Tanenhaus and Trueswell 2006). Because eye movements have been shown to be closely time-locked to spoken language (Tanenhaus et al 1995, Eberhard et al 1995), through eyetracking it is possible to see what entities that people consider as antecedents or referents as they hear sentences. Unlike experiments where only explicit responses and response times are evaluated, eye movements give us an insight into cognition in progress. In addition, the use of eyetracking methodology allows for relatively naturalistic experiments.

Participants may hear audio stimuli at a normal conversational pace, or even in actual

²Ordinarily in a DE context, if X **or** Y leads to Z, it is also true that X **and** Y leads to Z; the weaker item has entailed the stronger one. However, with the implicature, the assertion becomes "If X **or** Y leads to Z, but there is no information about whether X **and** Y leads to Z". Thus we know less about the circumstances leading to Z.

conversation, and in studies investigating reading, eyetracking has very little effect on natural language processing.

Although there have been eyetracking studies using reading on the topic of scalar implicature, even with specific regard to "some" (Bezuidenhout and Morris 2004), the present study uses the visual-world paradigm. In visual-world eyetracking studies, participants' eye movements are recorded as they look at a scene that represents the context of the sentences they hear, generally including intended referents. The scene may be an array of objects or entities, a real-world photograph, naturalistic drawing, or some type of composed scene in between. I am aware of only three published eyetracking studies directly related to scalar implicature in the visual-world paradigm (Storto and Tanenhaus 2005, Huang and Snedeker 2009, and Grodner, Klein, Carbary, and Tanenhaus 2010). (Important foundational research was also performed by Sedivy et al. (1999), which established that listeners are highly sensitive to scalar adjectives, such as "big" or "small", when objects on both ends of the scale are present.) It is worth describing these studies in detail as they are the first in this area, and ours will present an alternative approach.

Storto and Tanenhaus (2005), like Foppolo (2007) used the scale <and, or> to test listeners' early sensitivity to scalar implicature. In the case of "and", participants were shown an array of items that were both next to two identical instances of one kind of object or both next to two instances of two kinds of objects (Figure 1). Participants heard the sentence "The banana and the grapes are next to some locks. Please click on those locks."

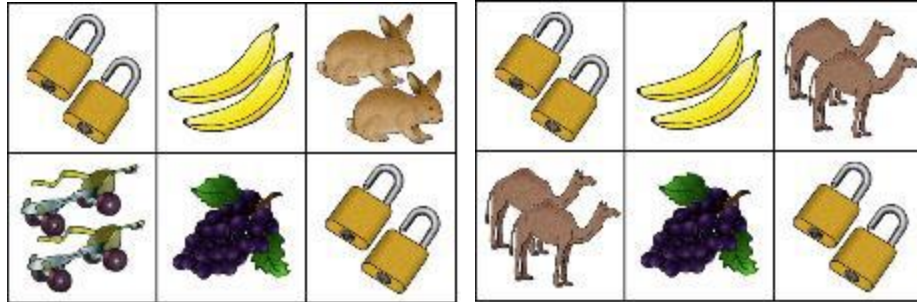


Figure 1. Storto and Tanenhaus (2005) sample scenes for "and". The left scene has one common pair of items, the locks, allowing for early disambiguation. The right scene has two common pairs of items, the locks and the camels, for the late disambiguation condition.

Participants were faster to look at one of the sets of locks when they were the only type of item the bananas and grapes had in common, indicating early processing of the meaning of "and". More relevant to the present study is "or", which, again, is claimed to have the implicature "not and". The experimenters compared scenes where objects were each next to an instance of one image and each next to a different image, or where the objects were all different (Figure 2). The stimuli were of a similar form: "The grapes or the oranges are next to some locks. Please click on the locks."



Figure 2. Storto and Tanenhaus (2005) sample scene for "or". The left scene has one common pair of items, the locks, for the early disambiguation condition. The right scene has no common items, for the late disambiguation condition.

When the objects had a next-to property in common, e.g. the skates, participants looked at the locks sooner than when the items were all different. However, they did not look at the locks before the onset of "locks." Thus while participants appear to process "not and" part of the meaning early, the authors claim that the fact that its processing was not as fast as looks to the target in the "and" condition indicates that a secondary process of costly inference may be at work, supporting the Relevance Theory view.

The other studies (Huang and Snedeker 2009, Grodner et al. 2010) are related to Storto and Tanenhaus (2005) in that they also involve a comparison between mini-scenes, but the scenes and the comparison are more complex. Huang and Snedeker (2009) aimed to detect how early people exhibited sensitivity to "all" relative to "some", as well as items on the scale of number terms, specifically "two" relative to "three". There were four mini-scenes, some with objects whose names began with the same sound; for example, a girl with 2 of 4 socks, a boy with the other 2 socks, and a girl with 3 of 3 soccer balls. (Figure 3) Participants were asked to "Point to the girl with some/two of the socks" or "Point to the girl with all/three of the soccer balls".

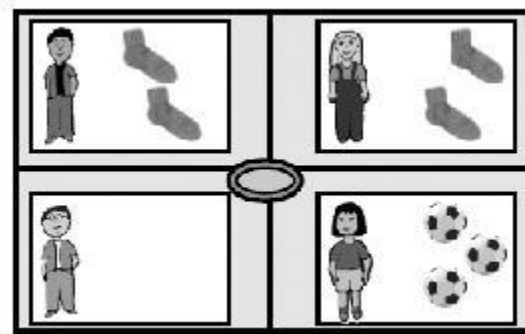


Figure 3. Huang and Snedeker (2009) sample visual scene.

Within 200 ms of hearing "all", "two", or "three", fixations on the target scene increased. As it takes 150-200 ms to program and execute an eye movement (Matin, Shao, and Boff 1993), this indicates that participants understood the meaning of these quantifiers quite early. This was not the case when hearing "some", for which fixations did not increase until the point of disambiguation (where "socks" is distinguished from "soccer balls"), but still before the end of the trial. Huang and Snedeker took this delay for "some" to indicate time taken to process the scalar implicature.³

In a directly related study, Grodner, Klein, Carbary, and Tanenhaus (2010) raise a criticism of Huang and Snedeker, namely that they did not show evidence of a literal phase before the "not all" inference was made as (Grodner et al. claim) would be expected in the Relevance Theory view. In other words, there was no increase in looks to the girl with all of the soccer balls, even though she also, literally, had some of the soccer balls. Among other modifications, Grodner et al added a competitor "none" condition, and found that the time to locate the target mini-scene was the same for "some" and "none". They conclude that the inference is made immediately on hearing "some" and does not contribute to processing cost, in line with the Default view.⁴

³Huang and Snedeker (2009) also showed that five-year-old children showed delay on hearing "some", but did not look at the target until the end of the trial, indicating lack of processing power.

⁴Although this chapter concerned only with adult processing, we should note that there are some which investigate children's capacity for scalar implicature (Papafragou and Musolino 2003, Noveck 2001), which generally suggest that this capacity is acquired late (though this may be an effect of experimental task; see Guasti, Chierchia, Crain, Foppolo, Gualmini and Meroni 2005 for an alternative view). Late acquisition is thought to be due to insufficiently developed processing capacity. As capacity is a factor, these tend to support the Relevance view.

Experiment

This experiment uses a different type of visual-world eyetracking methodology in hopes of giving a new perspective on the time course and required effort, if any, of processing scalar implicature. As previously discussed, eye movements potentially offer new understanding of the process of interpreting scalar implicature, but the ideal use(s) of eyetracking are being actively developed. The Storto and Tanenhaus (2005) approach, using an array of objects that do not combine to form a natural or meaningful scene, has the most precedent in visual-world eyetracking, and is most related to what the present work. However this type of scene is more often (though not exclusively) used when experimenters wish to know the participant's choice or anticipation of referent for some word, or the time course of that choice, according to a mental "linking hypothesis" of some image to some concept (Tanenhaus and Trueswell 2006). In the Storto and Tanenhaus experiment, participants were expected to discover and note relationships between objects, rather than find the object associated with a concept. It also seems uncertain whether the early looks to common objects in their "and" condition must be attributed to the common property, especially before participants knew what property was under consideration.

The Huang and Snedeker (2009) and Grodner et al. (2010) experiments use another type of visual stimuli, with four or more mini-scenes that must be conceptualized both separately and with reference to each other. For example Huang and Snedeker's participants saw Figure 3 and were asked to "point to the girl with some of the socks". In order to carry out the pointing, a participant must look at all the mini-scenes and create a

representation of which objects are distributed to which person, and whether this is a shared distribution or otherwise. The linking hypothesis that connects these complex representations with the mini-scenes is a step removed from that which links object concepts to object images. Additionally, Grodner et al. (2010) raises concerns about Huang and Snedeker's experiment with regard to variations in visual salience between the mini-scenes, as some have a greater number of objects than others.

The present experiment takes a different approach. Participants were not asked to choose the right picture from a set of pictures, but only to evaluate whether what they heard matched what they saw. Additionally, the experiment uses a visual environment that should require a relatively simple mental representation. Unlike visual-world paradigm eyetracking studies concerned with participants' choice of referents, our study is different in that all the objects are of the same type (e.g. apples). We do not expect looks to any particular object at any particular time, but rather that participants will note that there are a set of objects, all the same, or different in one property (color) only. As color and shape identification are low-level visual processes, and people are extremely fast to grasp the gist of scenes with considerably greater complexity than this one (see discussion in Henderson and Ferreira 2004), it is expected that participants will create a fast and sufficiently accurate representation of the scene.

As discussed in the next sections, this type of design allows us to evaluate different types of eyetracking measures than are normally used in visual-world eyetracking methodology. (Specifics of the experiment logic are introduced in the Predictions), While eye-movement patterns in this single-scene design will not reflect the

final interpretation of an entire utterance, as in the mini-scene studies, we hope that it will complement them by elucidating the processing that occurs during the utterance.

Methodology

Participants

Twenty-four students from the University of Southern California participated in the experiment and were paid \$10. All were adult native speakers of English reporting normal or corrected-to-normal vision including no colorblindness, and normal hearing. Data from two of the participants was excluded from the eye movement analysis due to inaccurate calibration of the eyetracker, but their explicitly-made task responses and response time data are included in the corresponding analyses. All participants were naive to the purpose of the experiment.

Materials and design

The experiment has a 2x2 within-subjects design. The two factors are Quantifier ("Some" or "All") and Picture (Match or NoMatch). The quantifiers in the scale <all, some> are used as sentences containing them are easy to depict, and as they are purely logical predicates, their interpretations are not affected by variations in participants' world knowledge.

The pictures are of two types. Either the picture has all the same objects in the same color, matching the quantifier, and is (Picture-All) or has two groups of objects that differ in color (Picture-Some), i.e. can be described by a sentence like "Some of the x are

y." If the picture agrees with the quantifier (QAPA, QSPS conditions), it is a Picture Match; otherwise (QAPS, QSPA) it is Picture NoMatch.

Each trial begins with a sentence naming the objects: "This is a picture of apples."

The second sentence is varied across conditions. The four conditions appear in Table 1.

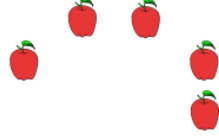
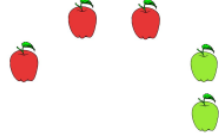
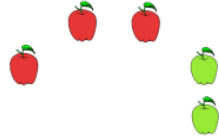
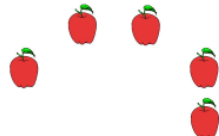
Condition	Example Sentences	Example Scene	Expected Answer
Quantifier-All-Picture-All (QAPA/Match)	This is a picture of apples. All of them are red.		Yes
Quantifier-All-Picture-Some (QAPS/NoMatch)	This is a picture of apples. All of them are red.		No
Quantifier-Some-Picture-Some (QSPS/Match)	This is a picture of apples. Some of them are red.		Yes
Quantifier-Some-Picture-All (QSPA/NoMatch)	This is a picture of apples. Some of them are red.		?

Table 1. Conditions, sample stimuli, and expected responses for the present experiment

Eye movements were recorded using an SR Research Eyelink II head-mounted eyetracker sampling at 500 Hz. The experiment was built and run using SR Research Experiment Builder software. There were 20 target items, 5 in each of the 4 conditions, as

well as 40 filler items for a total of 60. The items were presented in 4 pairs of lists, with each pair presenting target items in a forward and reverse order. Order of the items was randomized and fixed, and the items appeared in each of the 4 conditions within 5 experiment blocks, rotating the order of conditions across the lists (a Latin Square design).

The experiment was designed to elicit an approximately equal number of Yes and No responses, though an unexpectedly high number of Yes responses in the QSPA tipped the balance in that direction for some participants.

All images used in the experiment used instances of the same clip-art object in multiple positions within one of several fixed scene layouts. Each layout had space for 8 objects, 2 each at the top, bottom, left and right of the screen. For target items, there were two layouts, either of 5 or 7 instances of the object, with the remaining spaces in the layout empty. Each participant saw each layout an equal number of times. The objects in the picture were either all identical (Picture-All), or identical except for color (Picture-Some) (see Figure 4). In scenes where there was a color difference, the objects were presented in contiguous groups of two different colors. In this latter set, there were always 3 items that were the color named in the sentence, i.e. target items, while the other 2 or 4 objects were not. Thus the target items were the majority in the 5-object pictures and the minority in the 7-object pictures, in order to prevent comprehenders from equating "some" with "less than most". The colors were unambiguous and natural for the object (e.g., apples were red and green, not purple).

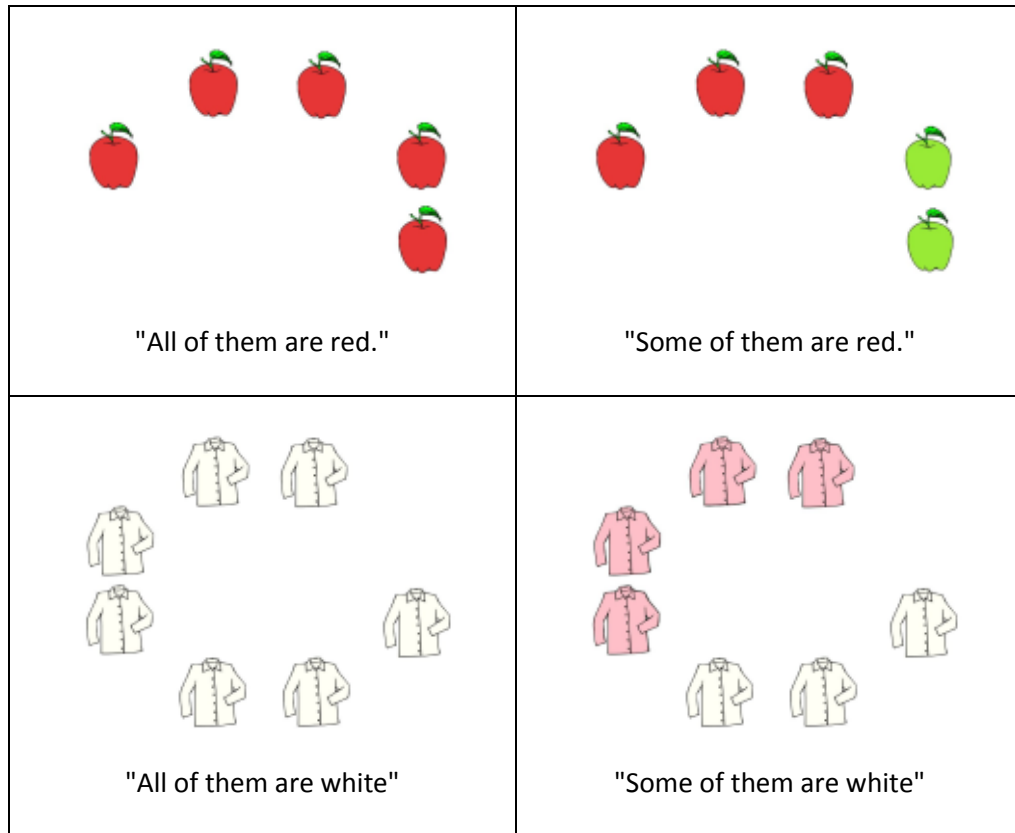


Figure 4. Sample stimuli for QAPA and QSPS conditions, in order to illustrate what we expect participants would call a "good description". The first sentence is "This is a picture of apples / shirts."

The filler items had their own set of fixed 8-position scene layouts, with the same object appearing in 1, 4, or 8 positions (Figure 5). Approximately half had the same style of being identical except for color. In some cases the colors used were less easily named (e.g. peach, magenta).

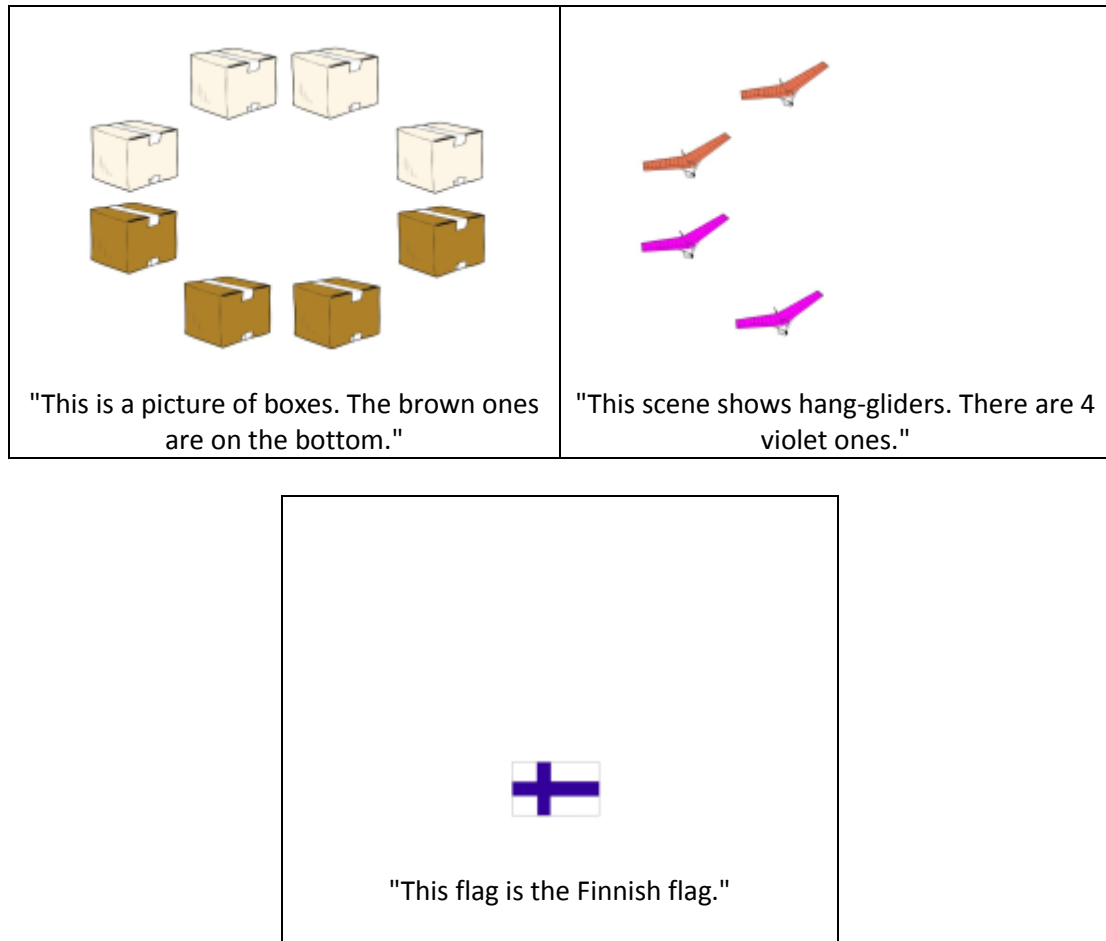


Figure 5. Sample filler images and sentences

Images were counterbalanced such that target and non-target items appeared in each of the 8 spaces in the layout an equal number of times.

Target item auditory stimuli consisted of pair of sentences, the first naming an object, the second a quantifier and the color of the objects. (See Figure 1 for complete list of conditions). For example, the first sentence might have been "This is a picture of apples". In the Quantifier-Some conditions, the second sentence was "Some of them are red", while in the Quantifier-All conditions it was "All of them are red". Filler item

stimuli also consisted of two sentences except for four one-sentence items. One-sentence items were of the form "This painting is the Mona Lisa", while two-sentence items varied but began with "This is a picture of boxes" or "This scene shows hang-gliders". The second sentence was usually a statement about color, but sometimes about the emotional state or location of the objects. No filler items were underinformative, e.g. they did not claim that four objects were blue when five were.

The names of objects were drawn from The Corpus of Contemporary American English (Davies 2008). Names of target item objects were among the most frequent nouns with regular plurals, and filler item objects were less frequent and not controlled for pluralization type.

Stimuli were recorded by a native English speaker in one recording session, using the software program Audacity and a Shure SM-58 microphone. Recordings were normalized to have consistent volume. In target items, the same audio recording of the first sentence ("This is a picture of apples") was used in both quantifier conditions of the item, and the second sentence ("Some/all of them are red"), recorded separately for each quantifier condition, was added to it, with 500 ms between sentences. Care was taken to avoid contrastive stress on the quantifier.

Procedure

Participants' eye movements were recorded as they looked at pictures and listened to sentences. The task for participants was picture verification. Participants were asked to listen to the sentence or sentences that accompanied each picture and decide whether the

sentences were a "good description" of the picture. Participants responded by pressing buttons on an Eyelink input unit, indicating "Yes" (good description) or "No" (not a good description). According to the instructions, if the sentences seemed wrong, misleading, or did not give enough information, participants might consider that a bad description, but they were to use their own best judgment.⁵

The complete experiment session lasted approximately 35 minutes. The time for the actual experiment, excluding the eyetracker setup and training and debriefing sessions, was approximately 10 minutes.

Data analysis

Response times: Response time data (in milliseconds) was trimmed to 3 standard deviations of each participant's response time, taking into account their times for both targets and fillers. This affected 0.8% of the target data.

Eye-movements: Successive fixations on the same object were counted as a single inspection of the object, with the durations of all fixations summed as the duration of the inspection. The start time of the inspection was the start time of the first fixation on the

⁵This task is not without concerns. Aside from issues related to verification procedures, which will be discussed later, a "good description" seems potentially vague (though the results showed that participants did interpret it with consistency). Still, there seem to be few other options. Other studies tend to ask for a truth-value judgment, which in our view too explicitly raises the issue of whether a "pragmatic interpretation" is really part of the definition of the word. This is something for the experimenter to consider, not the participant, and it may mask effects of generating the inference. In an alternative approach, Chierchia, Crain, Guasti, Gualmini and Meroni (2001) presented two sentences to children and asked which was better. Foppolo (2007) required a truth-value judgment, but added an interesting secondary measure to this by asking participants to rate on a 5-point scale "How much do you think the sentence is a good description of the situation represented in the pictures?" (127). We borrow Foppolo's idea but require a yes or no response.

object. The number of inspections combined in this way was approximately 71% of the number of raw fixations, both overall and on average for each subject. Fixations that were not on objects are not included in any analysis, and the number of these was consistent in all conditions. No inspections were less than 50 ms in duration.

Predictions

The logic of the experiment is to compare the hearer's response to "some" in scenes that make the "not all" inference relevant (Quantifier-Some-Picture-All condition), and scenes where it is not relevant (Quantifier-Some-Picture-Some condition). The difference in relevance is due to the informativity of the sentences. In the QSPS condition, "Some of them are red" is clearly a good description of a scene where some items are red, and the participant's response does not depend on a scalar inference, so the inference should be irrelevant. It creates no additional, useful information for the hearer, so according to Relevance Theory, the hearer should not bother to make it. In the QSPA condition, though, the sentence "Some of them are red" is underinformative relative to the picture. Since all of the items are identical, and there is no generalization that can be made about only some of them, the scalar inference should be uniquely relevant – informative, worth the hearer's effort – in this condition

Predictions about response times: Under the Relevance Theory view, if hearers do not bother to calculate an inference except in a context where it is important, then there should be indications of greater processing (longer response times) only for that context (Quantifier-Some-Picture-All). Outside of that context, results should pattern

with the Quantifier-All conditions, where there is no scalar implicature. However, on the Default view, if inferences from "some" are indeed so automatic as to be costless, there should be no difference in processing costs regardless of the quantifier. We would only be aware from participants' responses if they made the inference.

Predictions about eye movements: In addition to response times, eyetracking measures are also considered. In the Picture-Some conditions, there is a group of objects with one color, and a group of the same objects in another color. We report the time course of looks between the two color groups. We predict that when participants first come to perceive that "not all" is a salient inference, they will look to the other group, relative to the group they had been fixating previously. The time course of this shift should show at what point participants consider the "not all" inference to be salient

We also borrow some measures more often used in reading eyetracking studies, such as average fixation duration. It has been shown that the longer a person fixates a word, the more difficulty they are having processing it for one of many possible reasons, such as low frequency (Inhoff and Radach 1998). The assumption is that longer fixation duration on an object, or color-group of objects, is indicative of difficulty in general processing of the utterance. We also count fixations with the belief that more fixations, especially if repeated, demonstrate participants confirming for themselves that their initial perception of the scene was correct. This also has a parallel in eyetracking of reading, where refixations and regressive saccades are frequently used measures (Rayner 1998, Inhoff and Radach 1998, Clifton, Staub & Rayner 2007).

Results

We first look at participant responses and response times, which should point us in the direction of the processing effort, then turn to eye movement results for more details.

Yes/no Responses

Participants responded with great consistency, especially in conditions where the answer was clear without their needing to make any inference. Nearly 100% of responses to Quantifier-All-Picture-All and Quantifier-Some-Picture-Some items, also known as the Picture Match conditions, were "Yes", and to Quantifier-All-Picture-Some items "No".

(Figure 6)

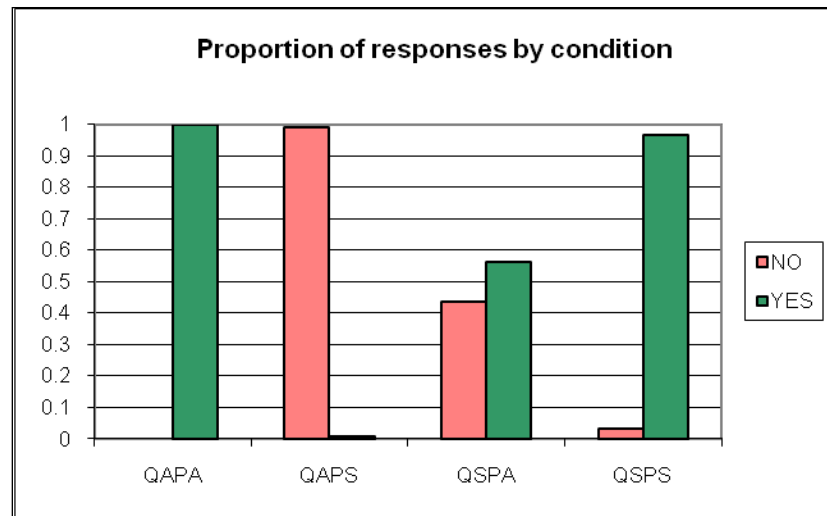


Figure 6. Participant responses.

Results for Quantifier-Some-Picture-All (QSPA) items were surprisingly mixed. The expectation was that underinformative sentences would not be considered to be a good description of a picture, especially given the explicit instruction that "not enough information" could be an example of a bad description. However, there were more logical-interpretation "Yes" responses (67 of 119 responses, 56%) than pragmatic-interpretation "No" responses (52 of 119 responses, 43%) to these items. There was consistency here also. 22 of the 24 participants gave the same answer in 4 or 5 out of 5 QSPA trials, with 9 consistently answering "No" (the pragmatic response) and 13 consistently answering "Yes" (the logical response). Such a split seems common, though in Noveck (2001) and Noveck and Posada (2003), more participants responded pragmatically than logically, the reverse of the findings here.

In both of those studies, as well as Foppolo (2007) where responses were also split (though she does not report participants' consistency), the experimenters interpreted the logical responses as evidence that some participants did not process the implicature. However, response does not seem to be a good indicator of this. In the present study, the experimenter directly confirmed awareness of the implicature in nearly all participants (22 of 24, 92%), either through the experiment responses (giving the pragmatic "No" response in the QSPA condition), or in debriefing after the experiment. Even participants who gave the logical "Yes" response were aware of the pragmatic interpretation, often volunteering it in an explanation for how they answered. We return to the implications of this issue in later sections.

Response times

A two-way repeated measures ANOVA was performed on the response time data, with the factors of Quantifier (Some/All) and Picture Match (Match / NoMatch to quantifier). Response times are the time from the end of the second sentence ("Some of them are red") to the time when the participant pressed the button. (Figure 7) There was no main effect of Picture Match. There was a main effect of Quantifier. Items in the Quantifier-Some conditions combined had significantly longer response times than items in the Quantifier-All conditions; when the picture was held constant, participants took longer to respond when the quantifier was "some". ($F(1,23) = 20.574, p < .01$; $F(1,19) = 10.327, p < .01$) There was also an interaction effect ($F(1,23) = 7.042, p < .05$; marginally significant by item: $F(1,19) = 4.163, p = .055$). There is a significant difference for both comparisons on the same picture. Paired T-tests show that QSPA was significantly slower than QAPA ($t(23) = -3.552, p < .01$; $t(19) = -3.259, p < .01$), and QSPS was significantly slower than QAPS ($t(23) = -2.933, p < .01, n.s. (p = .119)$ by item). In addition, participants took significantly longer to respond in the QSPA condition than the QSPS condition ($t(23) = -2.255, p < .05$; $t(19) = -2.114, p < .05$). On average, participants responded to the Quantifier-All conditions approximately equally quickly.

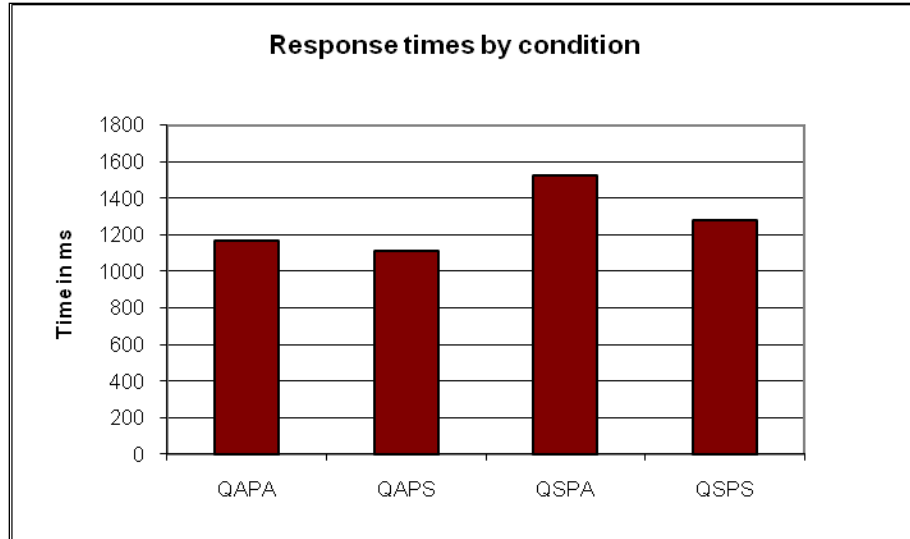


Figure 7. Participant responses.

Quantifier	Mean RT (ms)	SE (ms)
Some	1402	751.84
All	1137	443.52
Picture Match	Mean RT (ms)	SE (ms)
Match (QSPS, QAPA)	1193	522.63
NoMatch (QAPS, QSPA)	1345	714.75

Table 2. Participant mean response times by condition. The difference between Quantifier-Some and Quantifier-All is significant.

	Condition	Mean RT (ms)	SE (ms)
a.	QAPA (Match) 5	1165	475.43
b.	QAPS (NoMatch)	1109	408.92
c.	QSPS (Match)	1277	605.94
d.	QSPA (NoMatch)	1527	859.14

Table 3. Participant mean response times by condition. Significant differences are between (a) and (d) (the two PA conditions), (c) and (d) (the two QS conditions), and (b) and (c) (the two PS conditions).

As mentioned in the previous section, the QSPA condition was the only one where participants were split in their responses. Taking QSPA response time for participants grouped by their overall response tendency (including the one out-of-category response, for participants who were "4 out of 5" consistent) average response times are almost exactly equal: YES/logical group: 1519 ms; NO/pragmatic group 1502 ms. However, looking at QSPA response times solely by response, it appears that Yes/logical responses (1414 ms) were faster than No/pragmatic responses (1673 ms). An independent samples t-test confirms that this difference is significant ($p < .05$).⁶ This latter result is in agreement Noveck and Posada (2003), the only one of the studies where a direct comparison is possible; their logical participant group responded almost twice as fast as their pragmatic group.

In summary, in keeping with the assumption that longer response time is indicative of processing cost, we have overall found indications of cost for processing scalar implicature.⁷

⁶ However, the yes/no responses in the QSPA condition are not entirely independent, since, as previously mentioned, it occasionally happened that a person responded 4/5 times one way and 1/5 times the other. Since if we excluded these cases, there was not enough data to perform either an independent-sample or paired-sample t-test, this result should be regarded with some caution

⁷To mention some comparisons the reader may wonder about: there was no significant difference for response times when compared solely by response (participants were not faster to say yes), both in the case where target and filler responses were considered and when only target responses were considered. There was also no difference in response times for scenes with 5 objects compared to scenes with 7 objects, and no difference in the Picture-Some and Picture-All items (on their own, without regard to the quantifier).

Eye movement measures

Relatively longer response times are usually taken to indicate relatively greater degrees of cognitive processing, and a goal of this study was to provide an additional indication of processing through eye movements. Again, because the objects in the scene were all alike, except sometimes in color, for the most part we did not expect to find meaning in inspections of particular objects, but rather differences in overall patterns in how the scenes were evaluated.

In most measures, the time periods under consideration are different across condition. Most effects were found in the period beginning from the second half of the quantifier (the average length of the audio for "all" was approximately 200ms, and for "some" 300ms.) Thus the time period begins approximately 100ms after quantifier onset for the Quantifier-All conditions, 150ms after quantifier onset for the Quantifier-Some conditions), to the end of the trial. This time period was chosen because it is the earliest point where the quantifier could have had an effect on inspections (as participants needed to hear at least part of the quantifier word before initiating processing, and we assume that the midpoint is a reasonable approximation of that point), but as participants usually took longer to respond in Quantifier-Some conditions, the time period is usually longer for those trials. Since it is important to demonstrate that there is active processing going on that leads to this extra response time, and only secondarily interesting to demonstrate more active processing in the same time period across conditions, it seems relevant to compare these time periods even though they are different. However, there were indeed

cases where effects were found in fixed periods after the quantifier (see participant-contingent analyses).

Latency to switch between color groups

In the Picture-Some conditions, it is possible to see whether the distinction between the differently colored groups of objects is affected by the quantifier.⁸ Objects in the picture can be considered part of a target set (those with the color named in the sentence, e.g. the red apples when they heard "Some of them are red") or a non-target set (objects that did not have the stated color, e.g. the green apples). Participants were grouped based on whether they were inspecting target or non-target set items at the onset of the quantifier. Of course, participants were equally likely to be fixating either set at that point; the color was not named until (on average) 801 ms after the "All" onset, and 929 ms after the "Some" onset. Below is a plot of how soon the participant looked at the other group (the target group, if they had been looking at the non-target group at the quantifier onset, and vice versa) after hearing the quantifier (Figure 8).

Participants tended to look at the other group earlier in the QSPS condition than the QAPS condition. When participants heard "some", by 400 ms after the quantifier onset, they were equally likely to be fixating on the other group. However, when they heard "all", they were not equally likely to inspect the other group until 600 ms after the

⁸This comparison is not possible in the Picture-All conditions, because all the objects are identical.

quantifier onset (by item) or 800 ms (by subject). The same pattern of results was obtained when the comparison period was changed to the beginning of the phrase "of them".

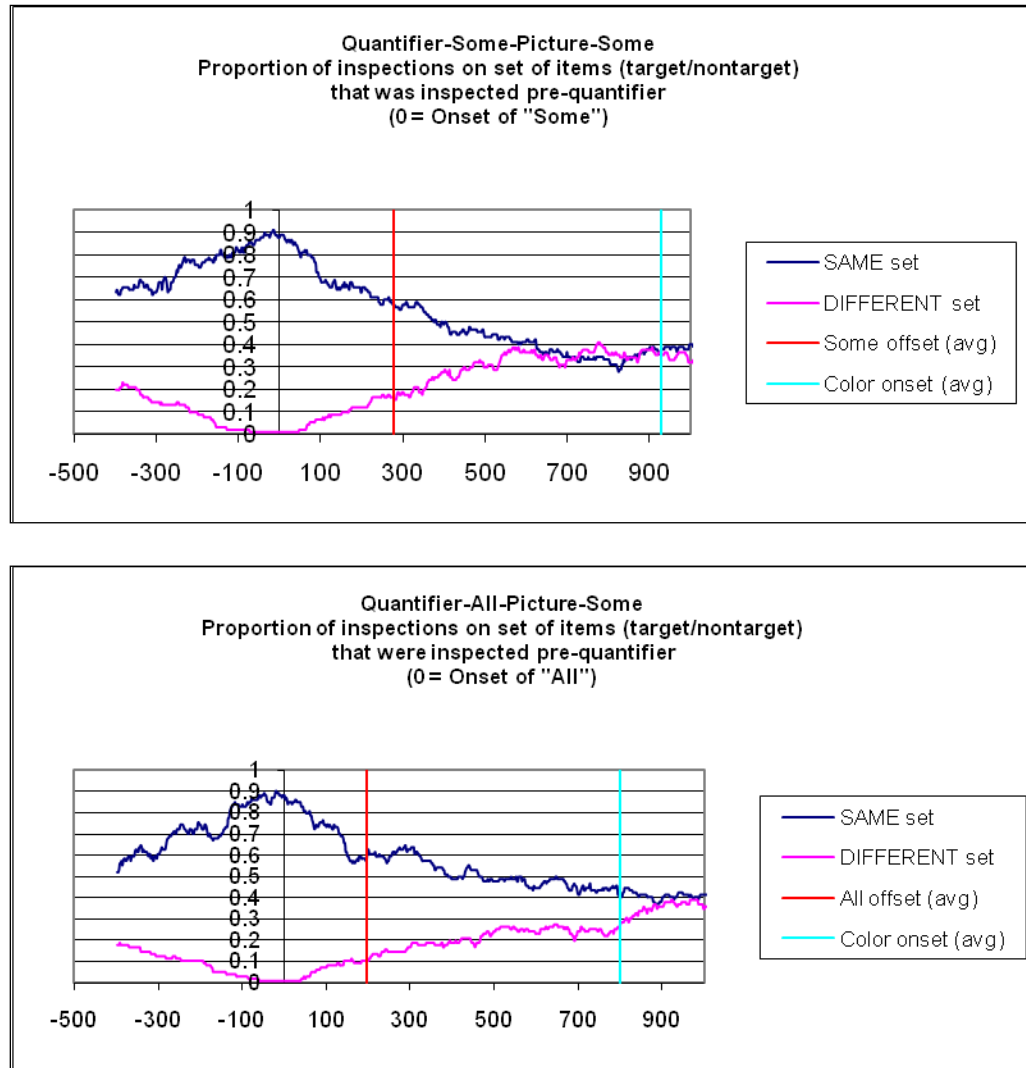


Figure 8. Proportion of inspections on set of items (target/nontarget) that was inspected pre-quantifier; zero is the point of the quantifier onset. For trials where participants were looking at the target group at the quantifier onset, looks to the target group thereafter are counted as "SAME", and looks to the non-target group are "DIFFERENT". If the participants were looking at the non-target group, looks to non-target group thereafter are included in "SAME", and looks to the target group are "DIFFERENT".

Participant-contingent analysis of eye-movements

Another example of early sensitivity to the quantifier can be seen by returning to the participant groupings by QSPA responses. Although both the QSPA Yes/logical and No/pragmatic groups had nearly equal response times, independent-groups T-tests show that the Yes/logical group had significantly more inspections (3.83) in the 200-600 ms period after the quantifier offset than the No/pragmatic group (2.43). ($t(16.878) = -2.389$, $p < .05$). Furthermore, inspections initiated in this time period led to an overall longer fixation time for the Yes/logical group (2141 ms) than the No/pragmatic group (926 ms) ($t(14.158) = -2.475$, $p < .05$) though there was no significant difference in average inspection times. These results appear to demonstrate greater processing for the logical group in the period following the quantifier. There were no differences in fixation count or fixation time when longer time periods were considered.

Average inspection duration

A two-way repeated measures ANOVA was performed on average inspection durations, with the factors of Quantifier (Some/All) and Picture Match (Match / NoMatch to quantifier). There was a main effect of Quantifier. Inspections of objects are on average longer in Quantifier-Some conditions (478 ms) than Quantifier-All conditions (417 ms), when counting object inspections that are initiated in the period from the second half of the quantifier to the end of the trial. ($F(1,21) = 5.120$ $p < .05$; marginally significant by item $F(1,19) = 4.032$ $p < .06$). There was no effect of Picture Match and no interaction

effect. When the time period was fixed in shorter periods, there was no significant difference in average inspection durations.

Summary of other eye-movement measures

We briefly looked at other eye-movement patterns typically used in eyetracking studies, such as total inspection time, re-inspections, et cetera, but these were not very informative. The remaining differences in eye movement patterns mostly correlated directly to the longer time participants considered the Quantifier-Some items. Specifically, there was a significant main effect of quantifier for the total inspection time, number of inspections (only significant by subjects), long inspections, and re-inspections (only significant by subjects), but only when the time period lasted until the end of the trial, which was longer in the Quantifier-Some conditions. That is, given a constant time period, e.g., 200-600 ms after the quantifier offset, means tended to be longer for Some conditions but were not significantly different.

Discussion

The data suggests that there is a consistent cost associated with "some", in agreement with most of the studies discussed previously (Noveck and Posada 2003, Bott and Noveck 2004, Storto and Tanenhaus 2005, Huang and Snedeker 2009). Longer response times and longer inspection durations both indicate greater processing in the Quantifier-

Some conditions. These aspects of the data seem problematic for the Default view, which proposes that scalar implicature processing should overall conserve effort. However, also like previous studies, it is not clear that it is possible to distinguish between the Default or Relevance views, since the costly inference was apparently made even in a context where it was not necessary for the task, and hence not worthwhile. This behavior is contrary to the predictions of Relevance Theory.

The Default view and Relevance Theory

Insofar as cost is associated with Relevance Theory (which we return to presently) one possibility is that the hearer considered the implicature sufficiently relevant in both Quantifier-Some contexts, and Quantifier-Some-Picture-All responses were slower for other reasons. We now revisit the results with an interpretation specific to this conclusion.

With regard to response times, at first glance, the data appears to support the Relevance Theory account. The Quantifier-Some conditions produce longer response times than the Quantifier-All conditions, indicating that the scalar implicature associated with "some" takes time to process. Furthermore, when the implicature would seem to be relevant in a picture context (that is, conflicting with it, in the Quantifier-Some-Picture-All condition), the mean response time is longest.

However, there is a lot more to be said about the interpretation of the response time results. The QSPA condition response times were longer than the QSPS response times, but this cannot be attributed to a difference in relevance between the two. This can

be concluded because the QSPS times response times were also significantly slower than the QAPS times. The hearer considered it worthwhile to process the implicature from "some" in the Picture-Some condition, as well as the Picture-All condition, and there is no reason to think that the inference should have taken longer depending on the picture type.

Why, then, are the QSPA response times longer, if not because the implicature was uniquely relevant there? The difference between the QSPS and QSPA conditions is the conflict between the picture and the implicature from the quantifier. The longer response times are thus possibly due to task-related factors. For example, it may be that participants were taking a long time to respond only during the first trial of an QSPA item, trying to work out the expectations were for a "good description" in this case. Having made the decision, they responded without delay in subsequent trials. This type of decision process was sometimes reported in debriefings.

This conclusion is borne out by the data. Response times for the first trial in the QSPA condition were indeed significantly longer than the average of subsequent QSPA trial RTs. Still, excluding the first QSPA trial from the comparison of response times in all conditions does not affect the significant differences between conditions that can be seen in Figure 2. Although there is a downward trend in response times for QSPA items over sequential trials, the significant difference between QSPA and QSPS response times does not disappear until we take only the fourth and fifth trials in each condition. Having made a decision, participants appear to get progressively faster at recalling it, but it takes a while. Crucially, though, even at this remove, there was still a main effect of quantifier; participants hearing "some" still took longer to respond than participants hearing "all".

Let us briefly consider an alternative view. Here, as intended in the experiment design, the context of the (matching) picture in the QSPS condition was not relevant enough to motivate processing of an implicature. The main support for this idea is that, although QSPS responses were significantly longer than QAPS responses, they were not significantly different from the QAPA condition; also the difference between QSPS and QAPS was significant by subject but not by item.

It is not worth going down this road further for the following reasons. First, QSPS and QAPS have the same picture; QAPA does not. It is not clear how reasonable it is to compare eye movements across the Picture-All and Picture-Some pictures, even though there was no a priori difference in eye movement patterns in these two picture types. Additionally, there were no significant eye movement differences found between the QSPS and QSPA conditions; although the response times were different, there is no other indication of different levels of processing. If we believe that QSPA response times were slow due to the processing of implicature, there is no data to exclude the possibility that QSPS times were slow for the same reason.

Eye movement evidence

Before discussing eye movements further, let us remind ourselves of the meaning of the measures. Longer inspection duration is usually associated with a greater consideration of the object being fixated, but interpretation here must be somewhat different. All the objects in the scene represented the same concept (e.g., all were apples). For this reason, we do not expect to find much variation in inspection durations triggered by particular

words. There was an overall longer average inspection duration in the Quantifier-Some conditions than the Quantifier-All conditions, but in order to achieve significance we must consider inspections for the entire second sentence ("Some of them are ...") to the end of the trial. Thus, two possibilities present themselves: (i) the longest inspections were towards the end of the trial (which was not the case) or (ii) each inspection was just a bit longer, indicating a slightly more complicated evaluation that is performed repeatedly, requiring more observations in order to achieve significance. This latter possibility seems more likely.

Eye movements also provide evidence that, as Breheny, Katsos, and Williams (2006) found, not only is the implicature processing costly, it is local, that is, calculated as the quantifier is heard, without needing to wait for the end of the sentence). In the Picture-Some pictures, participants were faster to look at the other group when they heard "some", and they did so well before they heard the name of the color. Early responses also confirmed another important aspect of our account, namely that participant responses cannot be relied on as an indication of whether the implicature was processed. Recall that we also saw early sensitivity to the quantifier in the QSPA condition, when the participant response groups were compared. The Yes/logical group initiated more inspections in the 200-600 ms after the quantifier (again, before the color word), compared to the No/pragmatic group, and those Yes/logical group inspections led to more inspection time. If we believed that answering logically meant that the implicature was not processed, this would not make any sense -- participants who answered pragmatically should have demonstrated more processing. The Yes/logical participants seem rather to demonstrating a greater and earlier awareness of the conflict. There were

no other indications of a spike in processing near the quantifier, or in similarly short time periods thereafter.

To summarize, we have seen evidence suggesting greater processing for "some". Since we always see it, though, regardless of the accompanying picture, should we say it is a demonstration of a Default, where the scalar implicature is *always* processed? Or that both the QSPS and QSPA conditions were contexts sufficiently relevant to trigger scalar inference? As this study joins many others in finding a cost, and it is crucial to the Default view that Generalized Conversational Implicatures (GCIs) be fast and costless, there seems to be greater support for the Relevance Theory account. We do not think these accounts wholly incompatible, though. Although processing an implicature appears to require a measurable inferential step, it seems reasonable that as the inference is required so frequently for GCIs, the links required for the scalar inference from "some" are strengthened to yield greater speed than other inferences. The delays here are not so long as to inhibit discourse; as we look at the later trials, the difference between responding to "some" and "all" is only around 150-200 ms.

This begs a question: if it were the case that the matching and non-matching pictures create a relevant context, in what visual context would the implicature be irrelevant? This may be a variation of what Tanenhaus and Trueswell (2005) call the *closed-set problem* of visual-world paradigm studies, where a limited set of referents presented in a scene creates an unnaturally restricted context (22). That is, our scenes may not be comparable to a linguistic context such as used in Breheny, Katsos, and Williams (2006). To consider a common example where it is argued that an implicature is not relevant:

(7) Q. Is there any evidence against them?

A. Some of their identity documents are forgeries. (Levinson 2000:51)

The answer is intended to mean, "At least some, I don't know about the others". But imagine an experiment rather like the present one. In one condition we present participants with a set of identity documents that vary in apparent validity, and in the other, the documents are all obviously and uniformly forged. If we asked these participants whether the answer given above was a good one, it seems likely that in both cases participants would indeed consider the implicature, no matter how they ultimately answered.

The closed-set problem is more applicable to the referential type of eyetracking, though. Tanenhaus and Trueswell suggest that the disparities might be partially due to having all the possible referents co-present rather than in memory (2005:26). Their main concern is that the pattern of lexicon activation in real-world language use is more diffuse than in visual-world eyetracking, making less than fully realistic use of lexical neighborhood effects. This is clearly a valid point, even in this non-referential study, as the scenes of the present experiment presumably activate the same concept repeatedly. However, it is possible that this is actually an advantage. Since fewer other concepts are activated, we see the activation effects of "some" with less interference. It may even be that, in Levinson's example above, that activation of the complicated concepts of identity documents, evidence, and forgery block the otherwise straightforward scalar inference. The chapters that follow provide more insight into this issue

Task effects

Another aspect of the closed-set problem is the potential for task-specific strategies. One possible solution is, rather than change the type of scene, is to use a task that does not involve verification, where participants would not feel obligated to be thorough. Storto and Tanenhaus (2007) offer some speculation on problems associated with a verification task, noting the confound of verification activity with sentence processing activity. Since verification activity is presumably held constant across all our conditions, we are less concerned with that issue than the fact that "an explicit verification task might encourage subjects to consider from the start interpretations that are not immediately considered in the normal processing of sentences." (Storto and Tanenhaus 2007:436). However, this is an issue in any task in any experiment. Also a verification task is useful in that it provides reason to look around the screen.

Still, there are several remaining concerns with the task. First, we asked whether it was a good description, not whether it was the best possible description, which may have caused participants to spend some time evaluating what a sufficiently good description was. We see some support for this in filler items. Participants responded uniformly for almost all of the filler items, and all filler response times were within 1 standard deviation of the mean response time for fillers ($M=1620$ ms, $SD=719.04$ ms). However, the two items with the longest average response times (Item 25: 2096 ms; Item 26: 2158 ms) did not have clearly correct answers, and were also the only two items for which there was substantial disagreement.

Another interesting consideration is whether or not it would be useful to give participants a “Not sure” option. If a substantial part of the delay in the QSPA condition was due to participants wondering what the experimenter considered a good description, allowing them the option to say, in effect, “it depends on if you mean ‘some and not all’” would reduce that time while not interfering with their generation of that inference, though response times might be longer overall. Also, it would allow participants the freedom to be unsure in other situations. (For instance, one participant took a very long time to respond to an item because she was not sure if the black horses were black or dark gray).

Some previous work has also manipulated the task by biasing participants towards a particular answering style, logical or pragmatic. In the first experiment in Bott and Noveck (2004), a variable in the experiment was instructions to the participants. They were either explicitly told to consider “some” to mean “some and possibly all” or that it should mean “some but not all”. Interestingly, participants seemed able to accept the logical instructions without difficulty, but not the pragmatic instructions. When asked to interpret “some” pragmatically, they were more often wrong (that is, answered “true” to “Some daffodils are flowers”, which should have been false because “some” meant “not all”, and all daffodils are flowers), and responded more slowly.

In one sense, it would have been preferable if participants had answered uniformly for the underinformative (QSPA) items. Although it seems clear that they were all eventually aware of the implicature, unless they rejected the description at the time, it is certainly possible that they might not have generated it for all trials for which they answered logically, or may have just been aware that something was off, but weren't

sure what the exact inference was until the debriefing.⁹ There was no way to get an on-line indication of whether the logical-group participants had made the inference, in part because it was not expected that anyone would accept items in the underinformative QSPA condition. The “latency to switch” graphs, showing that participants were faster to look at the “other” group of items when they heard “some”, are a helpful indicator, but not definitive, and not available in the underinformative QSPA condition.

On the other hand, it may ultimately be counterproductive to bias participants. The effect of such training would be to add a new piece of information to short term memory, that “some” *means* “not all”. If we did this, participants would reject underinformative items, but presumably with much less effort, and the results would not be distinguishable from the Default predictions. However, this might be worth trying in future experiment. If response times were still slower than in the “all” case, it would be strong support for Relevance Theory – adding additional information to the context has delayed processing. But if response times were similar in all conditions, it wouldn’t provide unequivocal evidence either way.

Conclusions

This chapter uses eye-tracking in combination with a single-scene verification task to examine whether the processing of scalar implicature presents cognitive difficulty for

⁹ In the debriefing, I asked participants first if any of the experimental items stood out. If they did not mention the QSPA items, I asked about “the ones with some”, getting more specific until I eventually described the some-and-not-all inference. Although it sometimes took multiple questions, all but one person mentioned the inference before I did.

hearers. Data from response time and in eye movement measures suggests consistently increased processing cost for participants when they heard "some" as opposed to "all" under comparable conditions, both in response time and in eye movement measures. We also showed that participants' responses to underinformative items may not be indicative of whether they were truly aware of scalar implicatures, but may rather show considerations of the task.

This chapter does not offer a definitive opinion on whether the Default view or Relevance Theory view is a more correct description of generalized conversational implicature, though the demonstration of cost tends to suggest the likelihood of Relevance Theory. In addition, the experiment was not designed to demonstrate any literal stage of interpretation preceding a pragmatic stage – it may be that the literal stage, if it exists, is too short to show in eyetracking. Incrementing the complexity of the visual context and/or devising a different task for participants may be the next step towards greater certainty on how scalar inferences are made.

Chapter 3: Lexical Decision Experiments

The first chapter provided an overview of the issues of defining a scale in scalar implicature. Words in scales may be ordered by entailment (a Horn scale, where items higher on the scale entail lower items) or non-entailing parameters such as temporal order (e.g., <succeed, try>, where succeeding tends to follow trying, though not this is not a logical necessity). Parallel to this debate, there has also been interest in the distinction between generalized and particularized scalar implicatures. The most substantial contribution to this has come from Hirschberg (1985), who argued for the theoretical unification of Generalized Conversational Implicatures (GCIs) and Particularized Conversational Implicatures (PCI). Hirschberg noted that something that intuitively feels like a scale can be generated by context (as in the example discussed in the first chapter, <speaking Greek, liking Greek food>) and therefore particularized (i.e. strongly context-dependent) scalar implicatures are possible. From that it follows that particularized scales are possible, with an order imposed on propositions or predicates by some salient feature in the context. If this is true, it begs the question of to what degree all scales exist due to context.

Word association and limitations on scales

Since the basic case for scalar implicature research is the lexical scale, it is arguable that the recency of mention of particular scale words, or more broadly, their association, could also be considered a type of context. "Association" refers to the underlying structure of

word relationships in the lexicon, which is dynamic to some degree based on the language user's experience, but large portions of which are more or less consistent over time across speakers of a language (Nelson, McEvoy & Schreiber 1998). Words are connected to each other in the lexicon by virtue of semantic, phonological or perhaps orthographic similarity, or because they tend to be used in the same situations, or both. These connections are referred to as associations. For example, the word "read" is associated with the word "book" both because both are related to the same subject matter, but also because the word "book" often follows the word "read". Similarly, "read" and "write" are associated due to their semantic relationship, and "once" and "upon" are associated in spite of the lack of a direct semantic relationship, because the words are often used in the sequence "once upon a time". Word associations can be demonstrated via a free association task, where a participant is given a word and then asked to produce the first word that comes to mind, as in the free association norms produced by Nelson et al (1998). It is also apparent in priming tasks, where participants will respond more quickly to a word that is more closely associated than one that is more distant (see discussion in Perea and Gotor 1997) .

On the one hand, then, the role of word association in scalar implicature seems worth addressing. Scales may not only be created from logical/semantic relationships, but also between words that become associated through use. It is then also necessary to turn the tables and consider the importance of the logical relationship. While it is true that the scales usually discussed with GCIs are ordered by entailment relation, there do exist sets of words that have this relation, but are not normally considered scales. For example, something that is "chilly" is also, necessarily, "cool". If something "constantly"

happens, then logically it also "sometimes" happens. Yet GCI research for the most part does not make reference to these other terms, in spite of the fact that they fulfill the property of entailment along with "cold" in <cold, cool> and "always" in <always, sometimes>.

Levinson (2000) follows Gazdar (1979) and others in describing scales as having an "aboutness" constraint. Words in a scale must be "about" the same idea, which is to say they should not individually have too many semantic features which are irreconcilable. In his view, this and other limits on scales are necessary in order to avoid overgeneralization of possible scales, or predicting scalar implicature where it does not occur. For example, "chilly" is often applied to air temperature or things that cause bodily cold, and "cool" is has no such specific constraint. Therefore an implicature cannot be drawn from scale <chilly, cool>, and that scale cannot exist.¹⁰

"Aboutness" seems more psychologically real when put in terms of word association. Closed-class words, or words like "cool" and "cold" that have fewer specific semantic features allowing them to be "about" a large, general set of topics, will be naturally more frequent. This is because as words have greater frequency, they are more associated with (connected to, interlinked with) other words through use. Thus frequent words are more likely to be connected to each other, and even more so if they also have semantic features in common and are usable in the same types of situations, e.g. to describe temperature. This leads to a further prediction that it is not merely that <chilly,

¹⁰ However it is not obvious how "aboutness" should affect Levinson's view automatic, default processing. If there is a fast link between "cold" and "cool", allowing hearers to process the negation of "cold" right away, why should there not also be such a link between "chilly" and "cool", if the entailment relation exists? The proposed answer is that such links themselves are weakened by the presence of additional semantic features. Perhaps also there is an intuition that there must be a limit on the number of such fast links, or else their existence would not provide much benefit.

cool> should not exist as a scale, due to the relative infrequency/semantic specificity of "chilly", but neither should something like <freezing, chilly>, as those words are each typically used with distinct topics, and less frequently, and are not strongly associated with each other or other words in the lexicon. The exception is when these terms have been recently compared, and thus become temporarily associated, as in (8).

(8) A: Is it freezing outside?

B: It's chilly.

To summarize so far: while entailment is not necessary for a scale, neither is it sufficient, and this can be explained via general word association¹¹ with little need for "aboutness" except as that can be used to explain association. More frequent words, which are most likely to be considered or chosen by a speaker, tend to be more connected to each other and other words in the lexicon. Furthermore, new connections formed by recent mention of words are especially strong, allowing for temporary scales as in <freezing, chilly> of the preceding example. These connections appear to play a role in forming scales.

Ultimately, I believe that if association is a primary factor, then this may well explain particularized scalar implicatures, such as those relating to Greekness. Before pursuing this line further, though, it is necessary to look deeper into the relationship word association and word scales.

¹¹ As previously mentioned, words may be associated with each other not only through their use in sequence, but also through some similarity, including semantic similarity and further including an entailment relationship. By "general association", or referring to association in contrast with entailment, I mean a connection between words in the lexicon that may exist for any of these reasons, but is known from norming to be particularly strong.

Association between scalar words

Word association has directional properties. For example, while a connection from "read" to "book" may be quite strong due to the typical ordering which those words are produced, the connection from "book" to "read", though it exists because of the shared semantic properties, is weaker likely because of the lack of reinforcement through word ordering. Typical GCI scales (that is, those that are most often examined in scalar implicature research) tend to contain more frequent words, which are due to their frequency are associated with many other words. Interestingly, however, my analysis of the data in the USF free association norms (Nelson, McEvoy & Schreiber 1998) shows the GCI scale words are often associated more strongly in the increasing order of the scale than decreasing. For example, in a scale like <cold, cool>, the higher word is "cold" because it entails the lower word "cool", but not the other way around ("cool" does not entail "cold".) Similarly, the association of the words "cold" and "cool" is stronger going from "cool" to "cold" than the other way around.

Table 4 shows the degree of association between some well-known scalar words. Participants in the USF free association study were given a word (the cue) and asked to write down the first word that came to mind (the response). The table gives the proportion of participants who gave a particular word as a response to the cue, demonstrating association between the cue and response. Let us compare the association between the lower scale word and the higher scale word that (in most cases) entails it. The top row of each pair of comparisons represents the proportion of cases where the higher word was produced in response to the lower word cue. The bottom row represents

the proportion where the lower word was produced in response to the higher word cue. Most important to observe is that the proportion in each top row is greater than in the bottom row.

	Cue	Response	Proportion
1	BELIEVE	KNOW	0.060
	KNOW	BELIEVE	0.011
2	COOL	COLD	0.108
	COLD	COOL	0.000
3	LIKE	LOVE	0.335
	LOVE	LIKE	0.044
4	SEARCH	FIND	0.491
	FIND	SEARCH	0.022
5	SIMILAR	SAME	0.414
	SAME	SIMILAR	0.047
6	SOME	ALL	0.086
	ALL	SOME	0.014
7	SOMETIMES	ALWAYS	0.372
	ALWAYS	SOMETIMES	0.028
8	START	FINISH	0.397
	FINISH	START	0.284
9	TWO	THREE	0.322
	THREE	TWO	0.088
10	TRY	SUCCEED	0.048
	SUCCEED	TRY	0.000
11	WARM	HOT	0.273
	HOT	WARM	0.034

Table 4. Word association proportions for common GCI scales, from data in the USF free association norms (Nelson, McEvoy & Schreiber 1998).

Table 4 thus shows that people are more likely to produce the higher scalar term in response to the lower term than the reverse, and that the order of these generalized scales is represented in unidirectional association. This may make sense for numbers, which

become associated through counting, and perhaps for terms with a typical temporal order, such as "search" and "find", but there is no obvious explanation for the remaining scalar terms. Note that the fourth item in Table 4, <find, search>, and the tenth, <succeed, try>, are not entailment scales, suggesting that association, however it has been formed, may be of greater importance than entailment in creating the scales for scalar implicature. Of course, the question of how these associations are formed is an important one, to be returned to in the discussion section.

This pattern, of lower scale words having a stronger directional association with higher words than in the reverse direction, is found for the typical GCI scale, but there is little or no association between other entailing terms: "chilly" is not an associate of "cool", according to the USF data. (This is not to say that the word "chilly" is not connected to "cool" in the lexicon in any way, but rather the connection spread across some distance or is otherwise weak. Again, "association" refers to *strong* association). The three experiments in this chapter manipulate the relation of entailment and association in order to see which is more important to scalar implicature. Ultimately, if it can be shown that association plays an important role at least in some cases, it provide new motivation for a unified account of generalized and particularized scalar implicature.

In the end, it appears that neither general association nor the specific relation of entailment can be isolated as factors in a scale, at least with regard to the pre-existing associations between words in the lexicon, as opposed to association created by the recent mention of some two words. Where differences between entailing and associated terms exist, we see the most evidence for implicature processing in the form of negation of higher scalar words that both entail and are associated with the lower ones.

Experiment 1: Entailing and Associate scalar words

If hearers understand speakers' scalar implicatures because of the entailment relation of scale words, we expect to see evidence for scalar implicature processing with any entailment scale of words. If overall association between scale words is the driving factor, we expect to see implicature processing for any scale containing words that are measurably associated, even in the absence of a semantic entailment relation between the words.

Experiment 1 manipulates the status of the higher word in a scale, the one that would be negated via scalar implicature. For example, for the common scale <cold, cool>, use of "cool" would negate the higher word "cold". Specifically, we manipulate whether that higher word entails the lower word (e.g. "chilly"), is an associate of it ("hot"), both ("cold"), or neither ("nice"). Participants heard the lower word ("cool") on the scale within a context sentence that was designed to elicit processing of the implicature. We measured whether hearer drew an inference from an apparent scalar implicature (i.e. processed the implicature) by comparing the response times for the higher words in scalar-implicature-supportive contexts to the response times for contexts that also contained the lower scalar word, but which were designed not to support implicature.

The technique for the experiment was borrowed from Macdonald and Just (1989) and Kaup (2001), who showed that words that had been negated were also inhibited, that is, harder and more time-consuming for hearers to access. At sufficiently long SOA (e.g., 1200 ms, but not for an SOA of 750 ms) Kaup's participants were slower to recognize a

noun that had been negated—specifically, when it had been asserted that the noun did not exist within the described situation.

- (9) a. Mary bakes bread but no cookies.
b. Elizabeth burned the letters but not the photographs.

In (9a), "cookies" yielded longer response times, as cookies were not present. However, in sentences like (9b), "photographs" did not yield longer response times, because they were present in the situation. Along with SOA manipulations, Kaup compared definiteness of the noun phrase, and whether it was part of an event of creation or destruction. Only the absence of the object referred to by the noun affected its accessibility.

Scalar implicature involves the negation of higher words on the scale because the hearer infers that the speaker means that the higher word does not apply. Therefore, it may be possible to tell when a scalar implicature has been processed for a particular higher scalar word when the participant takes *longer* to recognize that word. Therefore, the logic behind this experiment is to compare response times to target words that are higher scalar items of one sort or another. If a participant is slower to respond to any target words that entail the lower scalar prime, we would know that scalar implicature is affected by entailment relations. The same would hold true for associated target words.

There are certainly reasons for concern in comparing Macdonald and Just (1989), Kaup (2001) and preceding experiments, with the present experiment. The present experiment involves the negation of adjectives, verbs, and quantifiers, but not nouns, and

thus does not represent the absence of an object from a situation. However, consider why the names of negated, absent objects become inhibited. Presumably, as the objects are not present in the situation being reasoned about, it would not be desirable to keep them available and activated for future inferences. Activation of any inapplicable words could lead to errors in reasoning or language comprehension, so inhibition of such words may be something of a check against faulty analysis. It thus seems reasonable that this logic could also to the inapplicable stronger terms that are negated via scalar implicature. Just as it should be difficult to draw further inferences or associations from the non-existent cookies in (a), it should also be difficult to draw inferences or associations from false or inapplicable higher scalar terms.

Methodology

Participants

Participants (N=25) were adult native English speakers with normal or corrected-to-normal vision, who were members of the USC community. They were paid \$10. One additional participant was excluded after getting fewer than 70% of the comprehension questions correct.

Procedure

The experiment was designed and conducted with E-Prime software, developed by Psychology Software Tools. A CRT monitor was used for the most accurate recording of

response times. The experiment method was cross-modal lexical decision. Participants heard stimuli over headphones, with the exception of comprehension questions which were presented visually. After the items had played, a fixation cross appeared for 1200ms, and then a target word. The target word stayed on the screen until the participant responded.

Participants were instructed to pay close attention to the sentences they heard while keeping their eyes on the screen, and to answer whether the word they saw was a real word of English, or not a real word, by pressing one of two buttons on the E-Prime button box. The instructions included that they should try to respond as quickly but also as accurately as possible.

The entire session took about 30 minutes, including the instructions, a short practice session, and debriefing. The main experiment took approximately 10 minutes.

Materials and design

The experiment had a 2x2x2 design (Context type x Entailing target x Associate target), for a total of 8 conditions.

Two sets of sentential contexts were created, each consisting of 1-3 sentences, to contain the prime words. Those in first set, referred to as the Implicature contexts, were intended to evoke implicature processing. For example, an Implicature context for the lower scale term "sometimes" (usually thought of as part of the scale <always, sometimes>) is below.

(10) Many people think Jane is not reliable. You can count on her sometimes.

The intended meaning is that Jane cannot always be relied on. Contexts in the second set, referred to as the Basic contexts, were intended to block or at least not support implicature processing. Below is the corresponding example for "sometimes".

(11) Our co-workers think Barbara has no friends to go out with. I've seen her at Starbucks by herself, so she is out alone sometimes.

The speaker does not intend to say that Barbara is not always out alone. Rather, she is out alone sometimes, and possibly always.

Two Implicature contexts and two Basic contexts were developed for each of 12 prime words, to yield 24 in each condition, 48 total. The contexts were carefully normed, as discussed in the norming section below. Thus each participant heard each prime twice within the experiment, though separated as far as possible.

For the most part, the Implicature and Basic contexts using a particular prime word were developed as pairs, with one being slightly altered to support implicature. Thus the contexts had approximately same subject matter and same number of words. The prime words always appeared as the verb or post-verbally in the last sentence of the context.

The contexts were read aloud by the author, and were digitally recorded and edited using Audacity. Care was taken to avoid contrastive stress on the prime words.

Twelve prime words for which word association data was available (in Postman and Keppel 1970; Nelson, McEvoy, and Schreiber 1998) were chosen from among those most commonly discussed in scalar implicature literature. These primes represented a lower item on an entailment scale. For each of those scalar primes, four target words were chosen that could be negated as part of a scalar implicature. The targets were in one of four conditions, with examples given for the prime "sometimes".

Entailing and Associate (E/A): The target word entailed the prime, and was a strong associate of the prime. (Example: "always")

Entailing and Non-Associate (E/NA): The target word entailed the prime, but was not an associate of it, or was a very weak associate. ("constantly")

Non-Entailing and Associate (NE/A): The target word did not entail the prime, but was a strong associate of the prime. ("maybe")

Non-Entailing and Non-Associate (NE/NA): The target word neither entailed the prime nor was an associate of the prime. ("certainly")

The target words were in the same lexical category as often as possible (43 of 48 target words). Because there were very few words that could meet all the characteristics for each prime in each condition, it was not possible to control for word frequency when selecting targets. As might be expected, Associate target words were higher-frequency relative to Non-Associate words, and word frequency effects are evident in the final results. However, as will be presented in the next section, word frequency effects do not account for the entire pattern of results, and are not relevant to the overall conclusion.

There were 48 filler items interspersed between target items in the experiment for a 2:1 ratio with the targets. Like the targets, the fillers contained 1-3 sentences and had the approximately the same mean and standard deviation of word count as the target contexts. In order to achieve a balance of yes/no responses in the lexical decision task, 36 of the target words associated with the fillers were nonwords, obtained from the ARC Nonword Database (Rastle, Harrington & Coltheart 2002). The nonwords were pronounceable and obeyed English spelling constraints. The filler item target words and nonwords were normalized such that their average length was the same as for the target item words.

In addition, there were 21 yes-or-no comprehension questions designed to make sure the participants were paying attention, following both target and filler trials.

The stimuli were arranged in 8 lists. Each list was divided into 3 blocks, where each block contained an item in each of the 8 conditions, that is, a context in the Basic or Implicature condition, and then a target word in the E/A, E/NA, NE/A, or NE/NA conditions. The order of the conditions was initially randomized within each block, then the order was rotated within the blocks across the 8 lists. Each participant saw three target items in each condition. Because there were two contexts in each condition for each prime word (24 Implicature contexts, 24 Basic contexts, and 12 prime words), participants heard each prime twice within the experiment, but the lists were arranged these repetitions were spread across the experiment.

The complete list of stimuli appears in the Appendix B.

Norming of target stimuli

As the goal of the experiment was to compare how implicature processing affects different higher scalar item targets, it is necessary to make sure that the contexts were reliable in supporting or blocking implicature. In order to ensure this, the contexts were normed via a web-based survey. The survey was used to adjust and test the contexts until the Implicature contexts for a particular prime were consistently rated as more supportive of an implicature, with a given set of target words, than the Basic contexts for that prime with that same set of target words. In other words, the survey evaluated whether it was possible to contextually induce a scalar inference for particular E/A words or E/NA words corresponding to the prime¹².

Survey participants (N=136) were instructed to read the sentences that made up the context as if they had been spoken aloud. A question followed each context asking if the imaginary speaker “means or is trying to say, intends to say, or is implying” that one of the potential target words was not true or did not apply. Participants saw either the E/A word or E/NA word in the question (not both). They were told to rate the speaker’s intention on a scale of 1 to 7, where 1 indicated that they thought it was “Not at all likely” that the speaker intended to convey the given meaning (that is, the scalar implicature), 4 that the speaker may or may not have intended the meaning with equal probability on either side, and 7 that the speaker “Definitely” intended to convey that meaning. When

¹² Originally the survey question contained words in all four conditions, but it quickly became apparent that participants would never judge that the NE (Non-Entailing) words were in any way implicated by the speaker, because, for example, they were opposites of the prime, or completely unrelated to it. They were thus not used to norm the contexts. In addition, in several cases NEA words could not be found that made sense with the survey question about the speaker’s intentions, e.g. because the words were a different part of speech. Those words were used in the lexical decision experiments however.

there was a reasonable majority difference between the Basic and Implicature contexts with the same set of targets, that context and target set was used.

Data analysis

Participants' responses and response times to the target word were recorded and analyzed. Before further analysis, however, all responses to one target, *bump off*, which was the E/NA condition target of the prime *killed*, were discarded. There were 13 such observations, about 2% of the total number of all datapoints in critical conditions. This phrasal verb had been included because no suitable single word had been found that entailed *killed* but was not an associate of it, and which was also accepted in the norming process. However, because the participants had been instructed to answer whether or not they saw "a" real word, which many participants took to mean a single word, participants took very long time to respond and reported confusion in the debriefing.

Participants' response times to the target items were trimmed to 3 standard deviations from the participant's mean. This adjustment affected approximately 1.4% of remaining observations.

The rate of errors was extremely low for the lexical decision task, less than 1% for target items and 2.3% for filler items. On average, error rates for the comprehension questions were about 11%. No individual got less than 70% of the comprehension questions correct.

Results

Participants' sensitivity to the experimental manipulations was significant, but short-lasting. Figure 1 shows participants' reaction times to the first of each of the three trials in each condition. Significant effects were apparent when only the first trial was included in the analyses, and to some degree through the second trials, but not when the response times of these three trials were averaged to produce a mean for each participant in each condition. As we discuss in the next section, this supports the idea that the scalar implicature processing is subject to learning effects, and that repeated trials in an implicature-supportive condition will be misleading.

Considering first trials only, a repeated-measures ANOVA with factors Context type x Entailing Target x Associate Target showed a main effect of Context. Participants took significantly longer to respond to Implicature contexts ($M = 742.59$ ms, $SD = 55.65$ ms) than Basic contexts ($M = 670$ ms, $SD = 39.8$ ms), $F(1, 24) = 5.698$, $p = .025$. (Figure 9). There are no significant differences between individual target word conditions (that is, no Context x Entailing x Associate interaction), though the graph seems to suggest that the Context effect does not hold for the NE/NA words.

Note that the effect is significant by subjects, but not by items. We suggest that this is due to the large number of conditions, as response times that were averaged by item represented far fewer observations in each average than did those by subjects. Additionally, the most significant results were for observations of the first trial (or first and second trials) of each condition, reducing the number of observations even further.

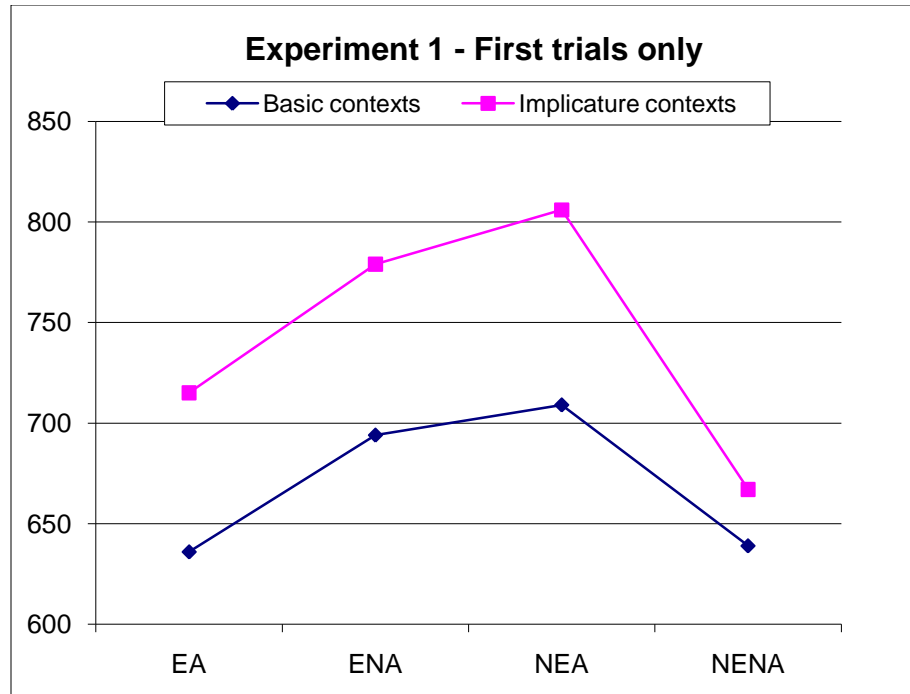


Figure 9. Response times (ms) in each of the target word conditions, separated by Context type. Only the first trial in each condition is considered.

We also looked at the RTs averaged across first and second trials, not just the first trial in each condition. This data is shown in Figure 10. The effect of Context appears to persist through the second trials. Considering first and second trials together, a repeated-measures ANOVA showed that Impicature contexts ($M = 738.45\text{ms}$, $SD=56.47\text{ms}$) are still slower than Basic contexts ($M = 690.98\text{ ms}$, $SD = 43.91\text{ms}$) (Figure 10) . However, the main effect of Context is no longer significant at the .05 level for a two-tailed test ($F(1, 24) = 2.976$, $p = .097$), though it would be as a one-tailed test. As the results are in the expected direction (Impicature contexts producing longer response times) as in many previous experiments, a one-tailed analysis of these results seems justifiable.

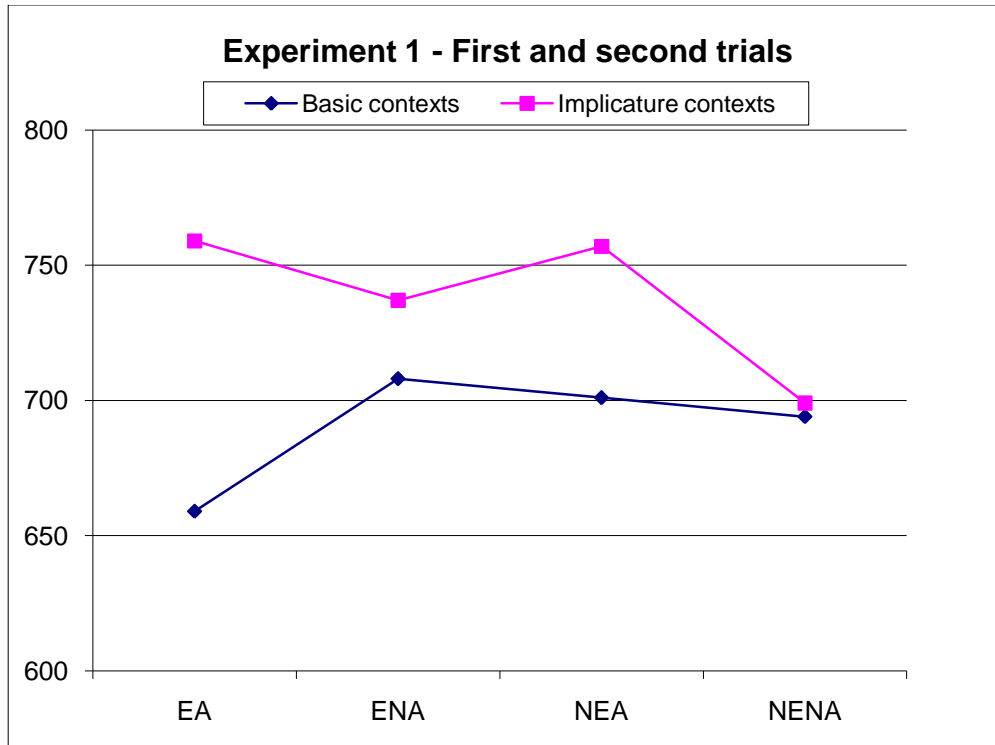


Figure 10. Response times (ms) in each of the target word conditions, separated by Context type, considering data from first and second trials in each condition.

In conclusion, the data suggests that context is effective in slowing response times for not only the typical GCI scale terms (the E/A target words, for example, "cold" in the scale <cold, cool> or "always" in the scale <always, sometimes>), but also for target words related by association or entailment individually, and not for target words that are unrelated to the prime.

It is interesting to briefly consider the set of second trials in each condition, that is, data from the middle third of the experiment. In that case, there was a (nearly) significant delay in responses to Entailing-Associate target words, but not to targets in any other condition. A paired-samples t-test showed that, taking only the second trials in each condition, Implicature context E/A condition response times ($M = 802.72$ ms, $SD =$

509.15ms) were slower than Basic context E/A response times ($M = 682.48$ ms, $SD = 289.03$ ms), $t(24) = -1.584$, $p = .063$ (one-tailed).

Effects of word frequency

Word frequency, expressed as occurrence per million words, was obtained for each target word from the SUBTLEXus database. (Brysbaert and New 2009), in order to find out to what degree frequency affected the response time results. A linear mixed model was created using subjects' repeated measures as a random factor, and target word frequency as a covariate. When looking at all trials for the experiment, regardless of any other experimental condition, word frequency was a significant contributor to RT, and there was a linear relationship between frequency and response time, where greater frequency correlates to faster RT. (Intercept = 692.69, estimate of frequency coefficient = -.009, $p = .058$). However, there were no significant effects of word frequency on RT for either the first or first and second trials. This was probably in part because the number of observations in these cases was insufficient for this type of model. Still, as results from the first trials indicate a *slowdown* in response time, even for frequent words (particularly those in the E/A and NE/A conditions), it appears that the effect of the experiment overcomes effects of word frequency.

Discussion of Experiment 1

The main result of Experiment 1 is that Implicature contexts produced overall longer response times than Basic contexts (Figure 9). There were no significant differences between target word conditions, and thus no statistical reason to believe that the Entailing or Associate status of the target word was a factor in the response time. If response times had been equally delayed for all target conditions in the Implicature context, this would indicate that the target word was not the source of the delay. Rather, it might have been that the processing of the implicature took place and spilled over through the 1200 ms SOA, slowing the response equally in all conditions.

However, looking at Figure 9, and keeping in mind that the relatively small number of observations considered (having excluded two-thirds of the collected data), we can propose a better explanation. It appears that the responses are uniformly delayed for all target word conditions except the NE/NA condition – where the target is unrelated to the prime. To put it another way, response times are slowed for all targets that *are* related to the prime, either through general association, or entailment specifically. Recall that the experiment was intended to demonstrate inhibition (reduced accessibility) of the word that had been negated via scalar inference. This interpretation suggests that, after a hearer processes a scalar implicature, any words related to the lower scalar term are inhibited.

Again, the support for the idea that negation leads to inhibition is indirect. The research that demonstrates it (Macdonald and Just 1989; Kaup 2001) applied to the negation of entities that were absent from context, not the negation of verbs, adjectives,

and quantifiers. We might also wonder why the RT delay is uniform given that the type of target-prime relationship (Entailing or Associate) is different. Are the connections between words equally strong regardless of whether their relationship is semantic or more basically associative?

Additionally, it is possible there is a special distinction to be made for Entailing-Associate targets. Along with the marginally significant effect of context for E/A targets in the second trials, Figure 10 also visually suggests a bigger difference between response times for E/A target words in the two context types, than other target word conditions. As the E/A words are the typical higher items in GCI scales, this is the expected result if the higher term were negated, and inhibited, via scalar implicature. E/A words were the overall most frequent, and are the most readily available terms for scalar implicature, and perhaps the strength of these connections means they cannot be inhibited so quickly. What is apparent overall, and important to note, is that the effect is opposite of what one would expect from high-frequency words. There is every reason that E/A target words should elicit the fastest responses, but the opposite is true. The process of negation and subsequent inhibition explains this.

While we cannot completely escape the possibility that implicature processing simply spills over through the SOA period (as there was no significant interaction between the target type factors and context), the evidence suggests that targets that are somehow related to the prime are inhibited, with perhaps the strongest inhibition applied to the E/A words, as would be expected in scalar implicature processing with the typical GCI scales. Non-associate words (E/NA targets) may also be negated via implicature, but as Figure 10 suggests, the effect appears to fade as early as the second trial in the

condition. Thus, we can tentatively say that scalar implicature does appear to apply mostly to terms are "about" the same thing or have the same level of semantic generality.

Although there was no significant interaction of Context and Association factors, it is likely that association alone (as in the NE/A targets) is responsible for some inhibition. However, it is important to note that this type of general (i.e. non-semantic) association, assumed from free association norms to be present as a baseline in the lexicon, may in reality be quite variable across participants (Nelson, McEvoy and Schreiber 1998). It is quite different from the patterns described in the introduction to this chapter, where words or predicates become associated from recent discourse. In Experiment 2, we address the importance of word association by applying contrastive stress to the prime words in the Implicature contexts.

Finally, it is certainly of interest that significant effect of Context is only present in early trials, specifically, the first trial in each condition of the experiment, which represents only the first third of the data. The experiment had been designed to take multiple measures for each participant in each condition, in order to get the most accurate picture of how the participant responds. As in Chapter 2, however, we find that repeated measurements in a condition can be misleading when it comes to scalar implicature. The context manipulation seems to be successful, but does not remain effective for very long. What exactly are participants doing that removes the effect? It is true that each prime word was used twice in the experiment, so having activated it and its associates once, the priming effect would be less apparent the second time. This does not seem likely though, because 1. the priming effect would not have lasted very long, and the uses of the

prime word were spread across the list, and 2. an analysis that looked at only the first instance of each item containing a particular prime word did not show any differences.

It is likely then that either participants come to recognize the Implicature contexts, or that the scalar inference itself, "not more" or "not [higher term]", was primed. This would be an important consideration for future experiments on scalar implicature. As we saw in Chapter 2, effort that is attributed to implicature processing might in fact be due to task-related factors that decrease over time (such as the participant's decision about what constitutes a good description). Similarly, apparent absence of processing, as here, might be related to a participant's very fast recognition of an implicature-supportive scenario. It will be important not to underestimate the hearer's fast ability to learn the patterns of scalar implicature, and check for such learning when evaluating data.

Experiment 2: Adding contrastive stress

Experiment 1 showed that active, frequent words were generally less accessible to the hearer after processing sentences in an implicature-supportive context. It also showed that there were no significant differences between target words that were generally associated with the prime, or target words which entailed the prime word. Given that an entailment relationship represents a probable, weak association to the prime even when it was not produced in free-association norming, Experiment 2 attempts to improve on the Experiment 1 results, or at least create a stronger overall effect of association, by using the same stimuli with contrastive stress on the prime word.

Contrastive stress is thought to draw attention to the words it is applied to, quickly bringing a set of comparison terms to the hearer's mind (Bolinger 1961, Rooth 1992). It stands to reason that these comparison terms must usually be strong associates of the original word, either recently established as such from discourse, or fundamentally within the lexicon of any speaker of the language. This is because, assuming that the attention has a constant effect on lexical access time, it would still take more time to retrieve the weaker, more distant associates than the strong ones. Strong associates of the stressed word would be available for contrast or comparison first.

The words evoked by stress are associates, but which subset of them makes up the comparison set depends on the context of the utterance (Rooth 1992). A speaker may be correcting a mishearing ("I said TALL, not small") or making a contrast between things in the discourse common ground ("Give me the TALL glass, not the short one."). If the context is scalar implicature, and the stress is applied to a lower scale word, the comparison term is likely to be the higher scalar word: "Jane SOMETIMES does her homework" i.e. not always. It is beyond the present scope to say how exactly a subset of associate words is selected for the comparison set, but the salient point is that associates are activated along the way.

In Experiment 2, we predict that Associate target words will pattern differently from Non-Associate targets, but as all the targets are linked to the prime in some sense (with the exception of the Non-Entailing/Non-Associate targets), there is a possibility for an overall inhibition effect. As stress should activate associates more strongly, then following the logic that negated words become inhibited, we should expect to find them inhibited more strongly if they are being negated via scalar implicature.

Methodology

The methodology of Experiment 2 was the same as Experiment 1, with the following differences.

Participants

Participants (N=24) were adult native English speakers as in Experiment 1, and were paid \$10. One additional participant was excluded for getting less than 70% of the comprehension questions correct.

Materials and design

The same contexts were used as in Experiment 1, including the same filler items and comprehension questions. However, the audio files for the Implicature contexts were re-recorded in their entirety with contrastive stress on the prime word (which, again, was the lower scalar term). The audio files for the Basic contexts were not re-recorded, and thus were exactly the same as in Experiment 1. This was because the addition of contrastive stress was clearly infelicitous in a non-implicature context, and for most of the items it was unclear what the speaker was trying to contrast. This would have led to misleadingly long response times, making the Basic contexts less useful as a baseline.

In addition, 21 of the 48 filler audio files were re-recorded with some instance of contrastive stress. In nearly all of these the stress was felicitous, with a few items left unclear to serve as extra-strong distractors.

The re-recording took place very soon after the original recording session for Experiment 1, using the same equipment, speaker, and settings. Although they were not normed, the new audio files were apparently indistinguishable in sound quality from those used in Experiment 1.

Data

As in Experiment 1, observations where the target was "bump off" were excluded. This affected fewer than 2% of observations. Participant response times were trimmed to 3 standard deviations of the participant's mean, which affected 0.9% of the remaining observations.

There were no errors on target items in the lexical decision task, and an error rate of 2.8% for filler items. On average, error rates for the comprehension questions were about 16%. No individual got less than 70% of the comprehension questions correct.

Results

We again conducted a repeated-measures ANOVA with factors Context type x Entailing Target x Associate Target. The results for Implicature and Basic context comparisons are

in the same direction as Experiment 1, though they do not achieve significance.

Implicature context RTs are slower ($M = 716.04$ ms, $SD = 35.95$ ms) than Basic context RTs ($M = 682.30$ ms, $SD = 39$ ms) ($F(1, 23) = 1.969$, $p = .172$) (Figure 11). This directional difference is present throughout. The Experiment 2 results also pattern like Experiment 1 in that significant effects are found when we analyze only the first of the three trials in each condition, and not when more data is taken into account. Learning or priming seems to have an effect very quickly.

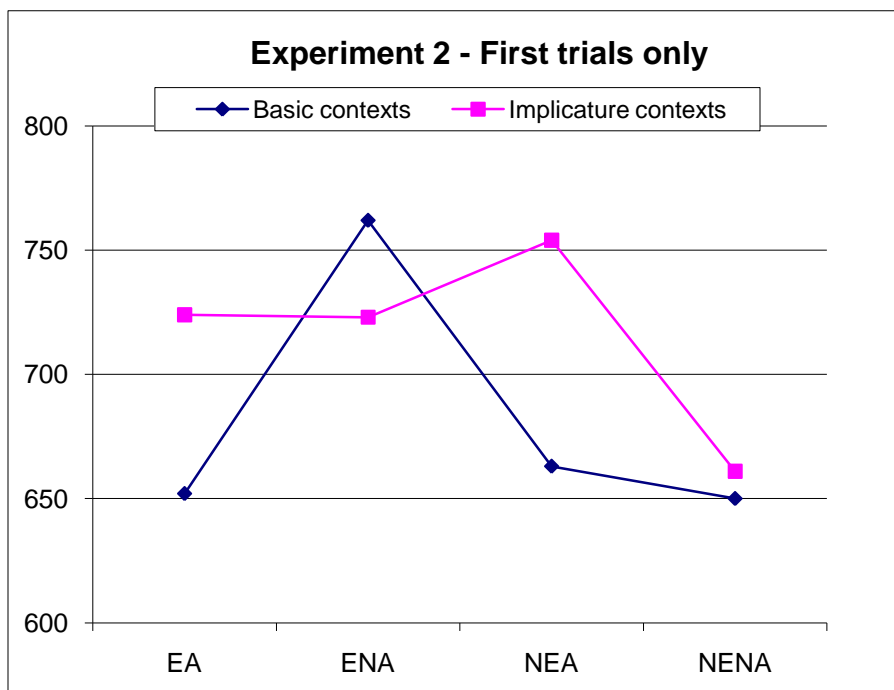


Figure 11. Experiment 2 response times (ms) in each of the target word conditions, separated by Context type, considering data from first trial in each condition.

However, Experiment 2 yields a new result in a marginal interaction of Context type and Associate target when considering first trials in each condition (repeated-measures ANOVA: $F(1, 23) = 4.093$, $p = .055$). A paired samples t-test showed that when the

target was an associate of the prime, participants were slower to respond in the Implicature context condition ($M = 739.17$ ms, $SD = 214.27$ ms) than the Basic context condition ($M = 658$ ms, $SD = 166.99$ ms) ($t(23) = 2.060$, $p = .051$). That is, the Implicature E/A and NE/A condition RTs, taken together, were slower than the Basic context E/A and NE/A condition RTs taken together.

Combining the results of Experiment 1 and 2 yields a stronger effect of Context type, where Implicature contexts are slower (repeated-measures ANOVA, taking into account Experiment as a between-subjects variable: $F(1, 47) = 7.444$, $p = .009$).

Analyses using first-and-second trial data, and the complete data from the experiment, did not yield any significant effects of interest. As the target words are the same throughout the experiments, the word frequency effects discussed in the Experiment 1 results section apply here also. Specifically, frequency contributes to the results in that more frequent words (generally those in the E/A and NE/A conditions) yield faster response times. However, while this effect is apparent when all the experiment data is taken together, the opposite is seen in the first-trial data. E/A and NE/A words have slower RTs.

Discussion of Experiment 2

Experiment 2, particularly in combination with data from Experiment 1, provides more evidence that contexts that support implicature are overall slower to process. In fact, the lack of a main Context effect in Experiment 2 also provides evidence against the idea that

implicature processing spilled over through the SOA period, delaying all response times uniformly. Rather, the target words themselves are the basis for RT differences.

More important is the apparent reduced accessibility of associated words after an implicature has been processed. Association should speed response times, not delay them, especially since it is positively correlated with overall word frequency. This result then also supports the tentative conclusion from Experiment 1 that, during the processing of scalar implicature, words somehow related to the lower scalar term are inhibited, in general. Furthermore, association, and not merely entailment, is a necessary property for a scale.

The fact that associates in general were inhibited, more so after they had been strongly activated via stress on the prime word, suggests that the same general mechanism of contrast may underlie both: TALL (not short), SOMETIMES (not always). The Non-Entailing / Associate words did not have any scalar relationship to the prime, and actually tended to be opposites, but were inhibited regardless. Yet when speakers say that something is cold, they are not implicating that it is not hot. The results of Experiment 2 may best argue for the general principle that negation can lead to inhibition of terms in any comparison set.

Experiment 3: Longer contexts

Relevance Theory (Sperber and Wilson 1995) states that hearers will make inferences, such as those in scalar implicature, as long as they have available cognitive processing capacity and the belief that the inference will be worth the effort. This has been

supported with regard to scalar implicature in De Neys and Schaeken (2007), where participants were less likely to process a scalar implicature if they also had to perform some additional cognition-intensive task. Experiment 3 uses this idea, attempting to tax participants' processing capacity by lengthening the contexts that they hear. Assuming that participants are, as instructed, listening closely to the sentences in anticipation of having to answer questions about them later, it stands to reason that the more material they must attend, the more their memory will be taxed, and possibly the more effortful inferences they will draw (though they may also try to conserve effort by trying to memorize the sentences instead). In any case, longer contexts seem reasonably to require greater effort for comprehension. In this experiment, then, we seek both to demonstrate both the effect that reducing processing capacity has on scalar implicature, and specifically whether the effect is most apparent on target words related by entailment or association.

The additional resource demands could manifest in response times in one of two ways. One is that implicatures will not be processed due to the additional demands, in which case there should be no difference in response times between the Basic and Implicature context conditions. Alternatively, if participants still do process the implicature, they should take longer to do so than in Experiment 1 due to the fact that they have more material in working memory. Further differences in target word types would continue to reflect differences in their accessibility.

Methodology

The materials and methods were the same as Experiment 1, with the following exceptions.

Participants

Participants (N=28) were adult native speakers of English, as in Experiment 1, and were paid \$10. No participants were excluded.

Materials and design

The materials were as in Experiment 1, but an additional 1-2 sentences were added at the beginning of each context to increase their overall length. The same material was used for each pair of Basic/Implicature contexts, and was therefore fairly general and designed to not interfere with the Context condition. For example, for one of the context pairs using the prime "sometimes", the Basic context was:

- (12) Barbara and Jane both lived in my building last year, below my apartment.
Our co-workers think Barbara has no friends to go out with. I've seen her
at Starbucks by herself, so she is out alone sometimes.

The Implicature context that was paired with that item was:

- (13) Barbara and Jane both lived in my building last year, below my apartment.
Many people think Jane is not reliable. You can count on her sometimes.

The preceding material was recorded separately some months after the audio for Experiment 1, and was prepended to the Experiment 1 audio files (which, again, did not contain stress on the prime word). Although the same recording setup was used, and the sound quality of the audio files was equalized as closely as possible, a few participants did report that the first part of the recordings seemed separate, even spoken by a different person. We will return to this issue in the discussion section.

An additional 1-2 sentences were recorded and prepended to 21 of the 48 filler items, so that the mean and standard deviation of word count in the contexts continued from previous experiments) to be equal across all target and filler items.

Data analysis

As in Experiment 1, observations where the target was "bump off" were excluded. This affected about 2.3% of observations. Participant response times were trimmed to 3 standard deviations from the participant mean, which affected 1.5% of the remaining observations.

The rate of errors was extremely low for the lexical decision task, less than 1% for target items and 3.2% for filler items. On average, error rates for the comprehension questions were about 12.1%. No individual got less than 70% of the comprehension questions correct.

Results

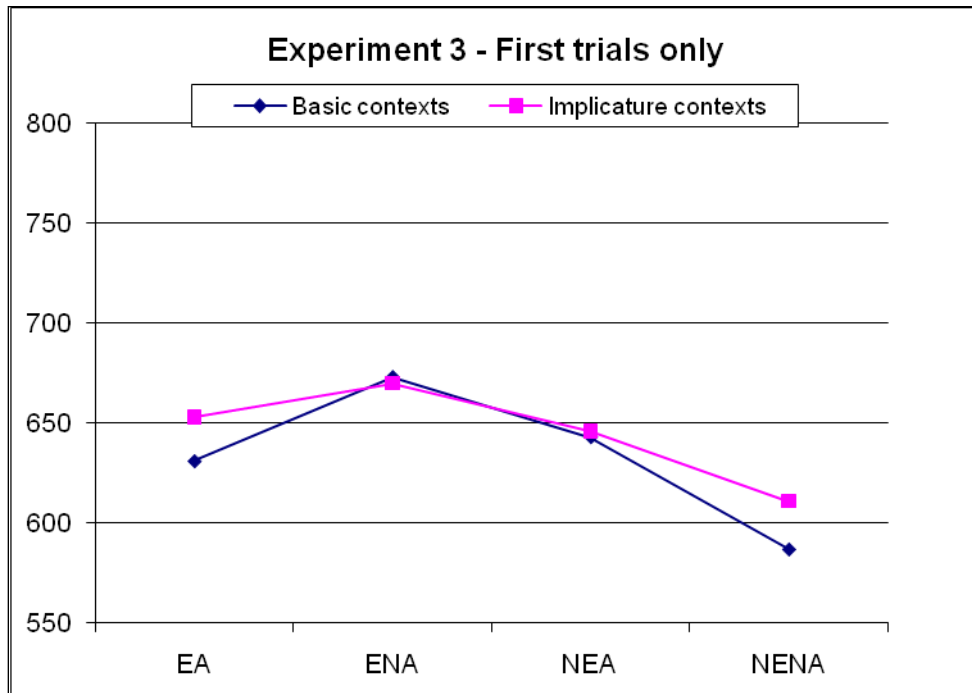


Figure 12. Experiment 3 response times (ms) in each of the target word conditions, separated by Context type, considering data from first trial in each condition.

We again conducted a repeated-measures ANOVA with factors Context type x Entailing Target x Associate Target. Following the pattern suggested by the previous experiments, only the first trials in each condition are shown here, but regardless of which trials were considered, there was no main effect of Context type, or any other expected effect. The only significant effect was a marginal interaction of Entailing target and Associate target ($F(1,26) = 3.865, p = .06$), where E/NA targets yielded slower response times than NE/NA targets. As this was not affected by Context type, it is likely attributable to word frequency in the absence of implicature processing.

Discussion of Experiment 3

It appears that having to attend to additional information reduces overall processing capacity, such that hearers either were not able to process the scalar implicature, or did not think the effort would be worthwhile. However, as the experimental manipulation essentially yielded a null result, this conclusion is subject to skepticism. We cannot discount the possibility that the additional content caused the Implicature contexts to lose their intrinsic implicature-supportive properties. In other words, it might not have been that having to attend to additional information presented too much processing difficulty, but that the actual content of the new information misled the hearer in such a way that the scalar inference no longer seemed appropriate. A future norming study will help to resolve this concern. Additionally, one wonders if the addition of a relatively small amount of material would induce that much greater of a processing load; after all, implicatures occur in the fullness of conversation. Therefore, it is possible that participants' processing ability was not overloaded, but rather just redirected as they attempted to make sense of the entire context. This might be even more likely if the additional material sounded different, as a few participants reported.

But there are reasons to doubt these ideas. First, the additional material, though recorded separately, didn't sound extremely different; the same computer, software, microphone, and speaker were used. 12 participants (43%) remarked on some sound quality differences in debriefing, but only after they were asked, and often added that it was not distracting. The quality difference within each context was consistent

throughout the experiment, and presumably the participants would have come to anticipate and get used to the variation.

Second, the additional sentences were designed to be as general as possible, and to have as little to do with the nature of the implicature as possible, while still making sense overall. It is true that excessive generality may have worked against them, causing the hearer to try to understand some apparent violation of the Relevance maxim (which would have been a confound if any difference between conditions had been found.) One or two participants did remark on how the speaker was saying random or unrelated things. But, as the overall context was experimental, the participants probably did not expect complete naturalness in the first place.

Finally, as previously remarked, implicatures do happen in full conversation. We have said that a little bit of additional content should not prevent the hearer from being able to process the implicature, because people obviously manage to understand implicatures in complex everyday discourse. However, in real conversation, hearers will not be attending each conversational utterance with the same effort. In other words, the difference was not merely that a few more words were added, but rather that there was a specific instruction, as part of the experiment, for the hearer to attend those words along with all the others in the sentences they heard. Therefore, it is quite possible that the greater attention needed for longer contexts induced additional cognitive effort that prevented implicature processing.

In summary, we cannot be completely sure that the additional material did not simply remove the essential difference between the Implicature and Basic contexts, but there are good reasons to believe that that was not the case. Rather, since participants

were explicitly instructed to extend the effort to attend the longer contexts, they were likely unable to process implicatures.

Conclusion

The results of Experiments 1 and 2 overall show that after hearing and processing a context that contains a scalar implicature, comprehenders are slower to respond to target words that are related to the lower scalar item from which the implicature stems. This indicates that the negative inference that hearers draw from the implicature makes these target words less accessible. We suggest that this is because they have been strategically (i.e. non-automatically) inhibited (McNamara 2005) via that negative inference (MacDonald and Just 1989, Kaup 2001). In fact, recalling that in early trials even the most frequent words were less accessible, we may go so far as to suggest that the more active a word is, the more strongly it will be suppressed via scalar implicature.

Crucially, the affected words apparently need not be higher scalar items in the sense of Horn scales, where there is an entailment relationship from higher to lower word. It is already well known that many lexical scales work very predictably in scalar implicature, even if the relationship between words is, for example, temporal, rather than entailment (e.g. <succeed, try>). The contribution of this chapter is to further specify why entailment is neither necessary nor sufficient for a scale. Words must generally be strongly connected to the lower scalar item in order to be affected by implicature. The strongest connections apparently also tend to have a semantic component (as in the Entailing/Associate words), but this does not seem to be necessary.

It is true that the nature of the required connection between words -- the source of their association -- can be hard to isolate. Words become associated through their meanings as well as through their use, and the Non-Entailing/Associate words all had some strong semantic relationship to the prime. Ultimately, though we seem able to say that any words linked to a lower scalar item, whatever the origin of the association, are affected by scalar implicature, not merely those that are logically stronger or entailing. If this is the case, then it seems likely that other context-based word scales, and even scales of full propositions, may serve equally well as input to the mechanism of scalar implicature.

As controversial as that idea has been, psycholinguistics research has not yet specifically investigated even the simpler issue of whether all Horn scales should be treated equivalently. It has often been the case that researchers doing experiments on scalar implicature have chosen a single scale, with the intention of using it to draw general conclusions on scalar implicature processing. There are relatively few results comparing scales within a single experiment. Papafragou and Musolino (2003) is one, comparing <all, some>, <three, two>, and <finish, start>; Huang and Snedeker (2009) and Grodner et al (2010) compare <all, some> and <three, two>; Breheny, Katsos, and Williams (2006) are concerned with <all, some> and exclusive and inclusive meanings of “or”, in experiments that were similar but not the same. Even in these cases, there is no detailed comparison between results of different scale types. The Huang and Snedeker and Grodner et al studies were in fact designed to demonstrate the already noted difference between numerical and other scales (a difference which was supported by Huang and Snedeker’s results.)

Among the many possible future directions for the research presented in this chapter, it would be very interesting to pursue scale differences further, especially given the new data on the importance of association between scale words. The relatively limited amount of data available from the present experiments has limited even the larger claims we can make, even before investigating the significance of the data on specific scales. As scalar implicature processing seems to be a fleeting, difficult-to measure phenomenon, even repeating the experiments with more participants is likely to yield more, and more interesting, results.

Chapter 4: Computational Model

In this chapter I introduce the Scalar Implicature Activation Model (SIAM), which shows that there need be no specific structure for scales in the lexicon, and that the associative links, and links between related concepts, account for the specific words that are most inhibited in scalar implicature. The results from Chapter 3 have suggested that scalar implicature seems to affect any words that are strongly connected to a lower scalar item, either via generally associative links in the lexicon or links that are semantic in origin. Crucially, it is not required that words be linked specifically via entailment, considered the classic case in Generalized Conversational Implicature research. In this chapter, we go a step further to show that no specific mental organization for scales needs to exist in the lexicon in order for a hearer to understand an implicature. This provides evidence that the different types of scalar implicature, such as those that do not involve entailment and those that do involve particular context, may all have the same underlying function cognitively, and should possibly be treated the same theoretically as well.

SIAM is a model of lexical comprehension, in that it reads a textual context and attempts to model how a hearer might go about processing (that is, understanding the meaning of) an implicature. It is based on human-generated data, from resources including WordNet (Fellbaum 1998) and the University of South Florida Free Association Norms (Nelson, McEvoy, and Schreiber 1998), as well as a framework based in the general connectionist principles often assumed in psycholinguistic models (to be discussed in the following sections). However, it should be kept in mind that its psychological reality is uncertain beyond these factors. Many assumptions about model

parameters, though relatively uncontroversial, have been freely made without considering additional psychological data. Still, within these limitations, the model achieves its primary goal of showing that scalar implicature is possible even without a formal structure for scales in the lexicon.

Overview

Computational modeling is often used in studies of language acquisition (e.g., Mintz 2003), to explicate the mechanisms behind phonological or morphological regularities (e.g. Cassidy, Kelly, and Sharoni 1999; Mirkovič, MacDonald, and Seidenberg 2005), or, like SIAM, to model behavioral data (see discussion in MacDonald and Seidenberg 2006). While many computational models of language make use of recurrent neural networks, SIAM is somewhat different in that while it incorporates many of the ideas of recurrent networks (such as nodes, feedback, and in some sense hidden layers), it is implemented in a somewhat more modern object-oriented framework.

A detailed description of SIAM's workflow will follow, but a general overview will make this easier to understand. SIAM is a connectionist model of a human lexicon, based on WordNet and other publicly available computational lexical resources. The nodes it connects are of two types: WordForms, corresponding to individual word strings regardless of what sense(s) they correspond to, and Concepts, which exist independently of words.

Links between nodes are directional: node A may be linked to node B, such that it is possible to spread activation from A to B but not vice versa. However, a WordForm

and the Concept it represents are reciprocally linked, modeling the fact that a person may wish to access a concept given a particular word, or choose a word based on a particular concept. A WordForm node may be linked to many Concepts (indicating polysemy), and a Concept may be linked to multiple WordForms (indicating synonymy).

Nodes may also be linked to nodes of the same type. For instance, the Concept node representing animals is linked to the Concept node representing dogs, indicating the semantic relationship of hyponymy. WordForms may also be linked, indicating word association independent of a semantic relationship. The links themselves are not distinguished from one another. A link that was created due a semantic relationship functions the same as a link created through any other type of relationship. (The reader may wish to skip ahead to Figure 13, which illustrates an example of these links and relationships.)

WordNet and the USF Free Association Norms (Nelson, McEvoy, and Schreiber 1998) contain the data for the links between nodes. Given a particular word, SIAM queries these two resources to find out what links to other concepts or words should be activated. During processing, nodes corresponding to these are created and added to SIAM's active node set, along with the nodes' baseline activation. Baseline activation for a node is calculated according to WordNet frequency counts. For Concept nodes, baseline activation corresponds to the \log^{13} of the sum of the usages of a word to refer to the given concept. To give an example, the baseline activation for the noun concept of water would be the count of instances that any word refers to the noun concept of water (e.g., "water", "H₂O"), as opposed to the verb concept of watering a plant. (This is not to

¹³ Log transformation was used simply to convert the word count data to a more usable scale, and does not represent any theoretical commitment beyond that.

say that the noun and verb concepts are unrelated, or that they would not activate each other. Rather, we are making the assumption that they are fundamentally separated, and thus have different levels of *baseline* activation.) For WordForm nodes, activation is computed by calculating the log of the summed frequency counts of each word in any usage or part of speech. The activation for the word "water" would include the count of instances it is used to refer to any noun or verb concept of water, watering, et cetera.

For this chapter, SIAM uses the contexts from the experiments in chapter 3 as input data, adjusting the activation of the words in those texts and the words that relate to them. Recall that the contexts were of two types: implicature-supportive and non-implicature-supportive ("Implicature" and "Basic" contexts, respectively, in terms of the experiments in chapter 3). It is important to note that SIAM cannot distinguish between these two context types, which is to say it does not "know" when an implicature is present. This would require drawing inferences on the basis of world knowledge, which no current system can do reliably in a general domain. To illustrate the problem, consider the implicature-supportive test item from the experiments in chapter 3: "You don't need anything heavy to wear. It's cool outside." The hearer is expected to understand that it is not cold outside. In part, that "not cold" inference is due to the awareness of the scale <cold, cool>, or (as this chapter suggests) at least awareness of the words related to "cool". However, world knowledge related to the context surrounding the scalar term, e.g., that people wear heavy clothing in cold weather, is of primary importance in a hearer's calculations, and is in fact the main difference between the Implicature and Basic contexts.

Because of the vast complexity and difficulty of world knowledge representation, SIAM does not try to draw knowledge-based inferences, and is thus obviously incomplete as a model of implicature processing. Rather, it treats scalar implicature as a purely lexical phenomenon, inhibiting the closest relatives of each word as they are read, in order to model the negative inference. Furthermore, the model applies inhibition/negation to every word it reads, rather than particular "scalar" words, which is unrealistic¹⁴. Thus SIAM should not be taken as an attempt to realistically model the whole of scalar implicature processing, but only to determine whether something like scalar implicature is possible when there is no specific lexical organization supporting a scale. The representation of scalar implicature that will be assumed is discussed in the model output section below.

Technical components

SIAM is an application written in the Java programming language. Fundamentally, it reads in a plain text file, "activates" an instance of a Java class for each node (i.e., instantiates a WordForm or Concept object, which are each derived from the Node object), storing the instances in a hashtable which organizes the current set of activated words and concepts. Each Node object stores its own hashtable containing references to parent nodes (which activated the current node) and children nodes (which were activated by the current node). Activation is spread by traversing the tree represented by these

¹⁴ SIAM was implemented this way both for simplicity, and because it seems possible that one might find other scale-like data in the output, besides those from the prime words. The output was so lengthy, however, that such analysis was beyond the present scope.

parent/children references. Finally, after each word in the text file has been processed, SIAM displays a text file representing the tree of nodes that have been activated for each word, and the activation level of each.

SIAM uses multiple publicly-available outside resources in constructing a lexicon and associating the words it reads to the appropriate senses. The SIAM lexicon is primarily based on the WordNet database, a resource which associates words to their senses or concepts. WordNet 2.1, the version used in SIAM, contains 155327 unique strings (which approximately correspond to words, though "words" containing spaces, such as phrasal verbs, are considered a unique string) and 117597 senses or concepts (referred to in WordNet terminology as synsets). WordNet also records the frequency with which a word is found in a particular sense from the many sources in the resource's creation. SIAM uses the log of this frequency count as a proxy for baseline node activation. Finally, SIAM's development was greatly eased by the Java API for WordNet Searching (JAWS) (Spell 2008).

The primary WordNet unit is the synset/sense/concept (henceforth referred to for brevity as concept). Actual words only exist in WordNet as they are linked to some concept. Therefore, searching on a word in WordNet brings up all the concepts associated with that word, and each concept also contains a list of synonym words for the concept. For example, searching on "water" brings up, among many other things, a verb concept that is glossed as "supply with water, as with channels or ditches or streams". This concept contains links to the words "water" and "irrigate". Thus the word "water" is directly linked to the concept of supplying water, and secondarily linked to the word

"irrigate". We can therefore imagine WordNet as representing a semantic network of interlinked words and concepts, and thus a loose approximation of the human lexicon.

The nature of the links of words and concept is somewhat limited within WordNet. Depending on the concept type (noun concept or verb concept, etc), concepts are linked to each other via hyponymy, hypernymy, part-whole relations, and a few more general relations such as "sisterhood" (terms that are commonly interchanged, such as colors) or "topic" (the general topic that might be under discussion when a particular word is used). The human lexicon has many more ways of associating concepts, both generalized (e.g., causation: "water" and "wet" are a plausible link) and specific to an individual, such as concepts associated with particular memories.

While some limitations must be accepted, it is at least possible to account for the links due to general word association. For this purpose, SIAM incorporates an additional resource, the University of South Florida Free Association Norms (Nelson, McEvoy and Schreiber 1998). The USF database was created by presenting participants with a cue word and asking them to write the first word that came to mind. A list of associated words were produced for 5019 cue words. In addition, the USF database contains the proportion of times a particular word was produced in response to a cue, which can be considered analogous to the strength of the link between them. While there are far fewer words in this database than WordNet, and their part of speech is not indicated, the USF data is a helpful supplement to the SIAM lexicon.

WordNet has other limitations that are problematic in representing a human lexicon. The database only contains nouns, verbs, adjectives and adverbs, and no function words such as pronouns or prepositions. A particular difficulty is the lack of

conjunctions, meaning that the scale <and, or> cannot be modeled without altering WordNet – something which is possible, but outside the present scope of work.

Helpfully, though, WordNet does contain some numbers, allowing the evaluation of <three, two> and a few other scales.

WordNet's concepts are organized by part of speech. It is not essential to have part of speech information when looking up a word. However, if the goal is to find the concepts most likely to be related to a word in context, the most appropriate subset of concepts will be returned if the word's part of speech is known. For instance, consider an input sentence "I run every day". When SIAM arrives at the word "run" and queries WordNet, many noun concepts will be returned, including "a row of unraveled stitches" (as in a stocking) along with the expected verb concepts. Of course many of the returned verb concepts will also be inapplicable (e.g., running for office). But to at least limit this type of misdirection, SIAM uses the Stanford Part of Speech (POS) Tagger (Toutanova and Manning 2000; Toutanova, Klein, Manning and Singer 2003) on input text prior to searching for it in WordNet.¹⁵ Each sentence in the input text is read as a unit, and then each word is tagged for part of speech and submitted individually to WordNet. This could be interpreted as modeling an ordered sequence of interpretation, where syntax precedes semantics, but that is not a commitment of this model. POS tagging is used here as a technical simplification to limit unrelated output.

¹⁵ This actually contrasts somewhat with research showing that comprehenders do briefly access multiple meanings, even across different parts of speech, when initially processing a word (e.g., Tanenhaus, Leiman and Seidenberg 1979) . However, given that these other meanings are quickly suppressed, SIAM approximates the effect by not activating them in the first place, and gains a processing advantage in doing so.

SIAM procedure

Although the technical details may seem considerable, the ordering of SIAM's processing is straightforward.

After initializing the data sources and POS tagger, SIAM reads in the specified file containing the text passage it is to process, identifying the sentences within it. Each sentence is divided into component words. Using the Stanford POS tagger, SIAM assigns a part of speech to each word.

SIAM then looks at the individual words. Let us say that the first such word is "chair". If a node representing the word string c-h-a-i-r (i.e. independent of its meaning or POS) is not found in the set of active nodes, a WordForm node for "chair" is created, activated with its full baseline activation (as it has only just been read), and stored in the set of active nodes. If "chair" had been previously read and was still active in SIAM, the WordForm node that corresponds to it is retrieved, and that node's activation is increased to the full baseline activation.

Next, SIAM spreads activation to words and concepts that are related to the just-activated word, "chair" (see Figure 13). Using the POS information about "chair" – imagine that it has been tagged as a noun, for now – SIAM queries WordNet for concepts linked to "chair", and creates and activates these as Concept nodes. (By "activate", we mean compute the appropriate level of activation for the node, and store it in the set of active nodes. If the WordForm or Concept node was already in the active node set, its activation is increased based on the just-computed value.) Returning to the example with "chair", we see that glosses for two such concepts are "a seat for one person, with a

support for the back" and "person who presides at the meetings of an organization". Next, SIAM queries the USF free-associate data for associates of "chair", yielding (for example) "table" and "sofa". These are created as WordForms, and again created and activated. SIAM then adds a link from the original WordForm node for "chair" to these new WordForm and Concepts nodes.

Because these new nodes are one step removed from the original, their activation is reduced proportionally, using the spread *degradation setting*.¹⁶ This setting represents an assumption that something like activation degradation or decay must occur as activated nodes grow further from the source of activation. The data reported here uses a spread degradation setting of 0.9. Nodes one step removed from the original word have their baseline activation multiplied by 0.9^1 (which is 0.9, decreasing it to 90% of the original activation); nodes two steps removed have their baseline activation multiplied by 0.9^2 (which is 0.81, 81% of the original activation), and so on.

This stage repeats for the number of iterations defined by the user (the *activation spread setting*), allowing SIAM to spread activation from each word to the desired depth. A spread setting of 1 only activates nodes directly related to the current node (what might be called *child nodes*). A setting of 2 activates child nodes and nodes directly related to them (*grandchild nodes*), 3 to spread to great-grandchild nodes, and so on. Again, the activation of these nodes is reduced proportionally as they are more distant from the current node.

¹⁶ All thresholds, proportions, etc. are definable by any user of the model, but a full exploration of their effects is outside the scope of this chapter. The settings used represent middle-of-the-road values that seemed to work reasonably well.

New nodes are found by looking at each node in the set created by the previous step. Again, the set of child nodes from the original word "chair" include the WordForm nodes "table" and "sofa" and Concepts "a seat for one person..." and "person who presides..." If the spread setting is at least 2, grandchild nodes are found by querying the lexical resources for:

- a. WordForms associated with the given Concept node
(e.g., other words used for "person who presides..." like "chairperson".)
- b. Concepts related to the given Concept node
(e.g., "a seat for one person..." relates to "furniture that is designed for sitting on")
- c. WordForms associated with a given WordForm.
(e.g., "sofa" is associated with "couch")
- d. Concepts associated with a given WordForm.
(e.g., the word "sofa" is linked to the concept "an upholstered seat for more than one person")

The newly active nodes will likely yield links to each other as activation spreads. Loops of links are often created, but activation is not allowed to spread infinitely across these links.

Activation of a given node may change positively or negatively. If an active node is reactivated directly or indirectly, it receives a boost in activation proportional to the activation of the higher (parent, grandparent, etc) node. Activation also degrades over time. The time where one word is read and activation spreads from it is referred to as a *cycle*. After each cycle, the activation of all currently nodes decreases uniformly by a

user-definable proportion referred to as *cyclic degradation setting*, which for the data reported here was 0.9. When each node's activation falls below a certain level, the *activation threshold*, it is removed from the set of active nodes. The activation threshold here was 0.1, roughly corresponding to words/concepts that only appeared once in WordNet's corpora¹⁷; a word must have appeared at least twice to be active in SIAM. Similarly, if the baseline activation of a node was below the activation threshold, it was not activated in the first place.

Finally, at the end of each cycle, SIAM models the effects of a negative inference on the current word.¹⁸ Again, the experiments in Chapter 3 supported the findings of Kaup (2001) in that words that were negated via scalar implicature became less accessible; hearers were slower to respond to them when presented within a lexical decision task. Within SIAM, this is modeled by inhibiting, or reducing the activation of nodes linked from the current node, similar to activation spread. The amount of inhibition is reduced proportionally as the nodes are further from the current node. The difference between inhibition spread and activation spread is that the proportion, or *inhibition degradation setting*, is much lower than the spread degradation setting, here 0.5, and applied less severely the further SIAM goes from the initial WordForm node. That is, nodes one step removed from the original word (child nodes) have their

¹⁷ WordNet does not report the overall word count of the corpora that were used to create it, so it is not possible to say that this activation threshold corresponds to, for example, 1 occurrence in 1 million words.

¹⁸ Each word is thus "negated", regardless of whether it is a scalar term. It is of course not realistic for a person to inhibit the nodes linked from every word that has been read. As discussed in the previous section, however, it is not easily possible to model the world knowledge that a human would use to decide which words merit scalar or other inference. We therefore limit the scope of SIAM to identifying what particular words would be inhibited as part of the SI process.

activation reduced by 0.5^1 (50%), nodes two steps removed (grandchild nodes) have their activation reduced by 0.5^2 (25%), and so on.

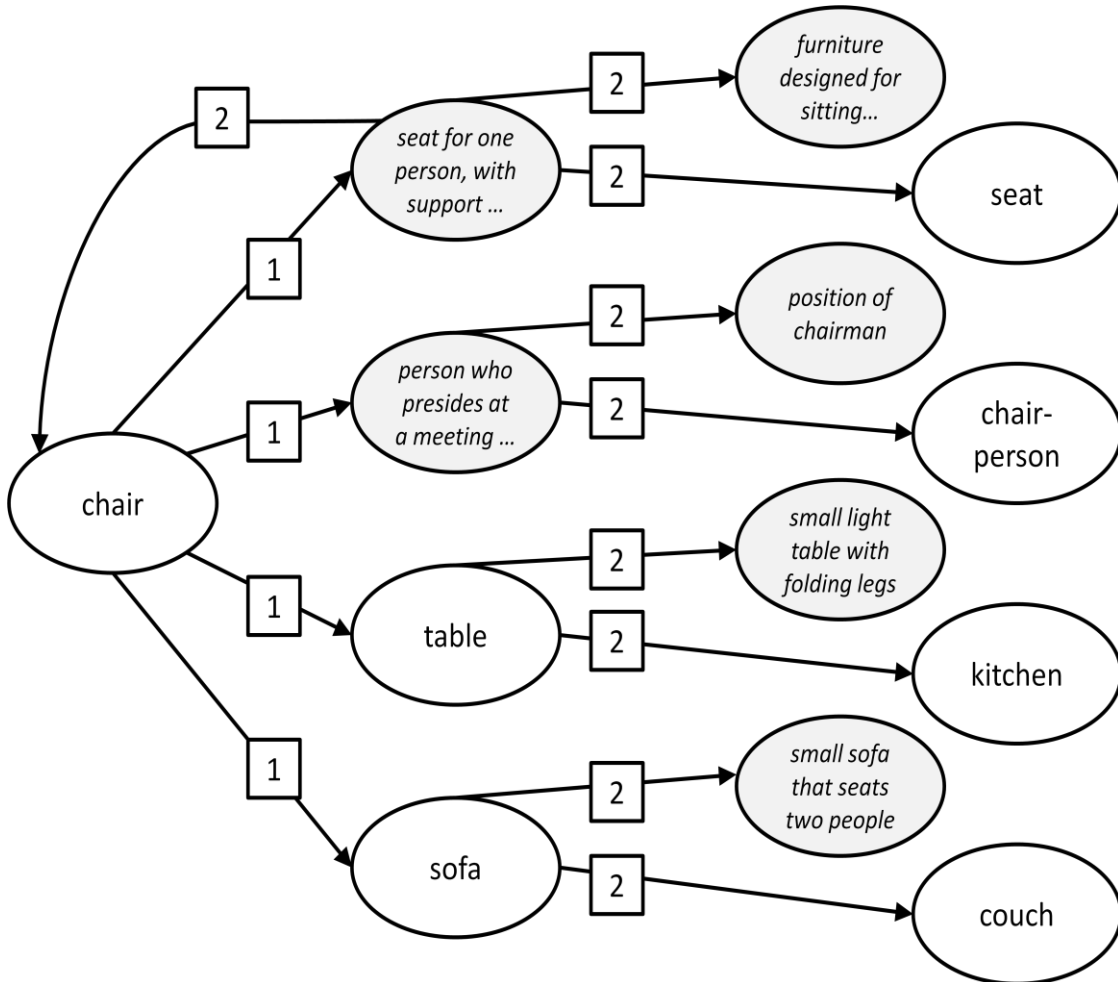


Figure 13. This diagram shows how SIAM spreads activation to new nodes. White nodes are WordForms, and shaded nodes represent Concepts. The nodes that are attached to links marked 1 are (a subset of) what would be created in the first iteration of activation spread. Nodes connected with 2 would be activated in the second iteration, and so on. Starting at the second iteration, reciprocal links (such as from the Concept "seat for one person" back to the WordForm "chair") would likely be added throughout, as SIAM identifies that concepts and words have a relationship in both directions. For simplicity only one such link is shown. Additionally, there would certainly be other links between the nodes present, and to nodes activated from other words in the text.

After all the sentences in the input text have been read, SIAM prints out each word in the text it read, and the list of all nodes that are linked at whatever distance to that word, along with the nodes' current activation level.

Model output results

SIAM read and evaluated each of the test contexts from the reaction-time experiments presented in Chapter 3, with the exception of those using "or" as a scalar prime¹⁹. There were a total of 44 short text passages, two each in the implicature-supportive ("Implicature") and non-implicature-supportive ("Basic") conditions for each of 11 prime words from Chapter 3. Within SIAM, no particular difference would be expected between the Implicature and Basic contexts, as the implicature-supportiveness derives from world knowledge (see previous discussion). SIAM only evaluates whether the inference is possible given a certain organization of the lexicon.

We evaluate which words are affected by a negative inference by examining the set of word nodes that were activated by SIAM's processing. In particular, we look at word node activation levels. Words nodes that are linked to a prime word (i.e. the lower scalar item), and that have a low activation level relative to other linked nodes but still above threshold, are assumed to be possible candidates for a scalar implicature stemming from that prime word. If the words were inhibited to below the activation threshold, we assume they would not be perceived as implicature – that is, even if the same mechanism had reduced their activation, a person would not be conscious of applying the inference to

¹⁹ As previously mentioned, this was because the WordNet database does not contain function words, with the exception of some numbers.

them. In other words, people recall thinking that "cool" implicates "not cold", because "cold" is still somewhat activated, but not "not chilly".²⁰ In other words, the central results of the study concerns nodes that should be highly activated (due to multiple links and high baseline activation) but are not, and yet are not inhibited beyond detection either.

Before getting into the specifics of the activated nodes, it is worth looking at the overall activation patterns. As the focus here is on word inhibition, concept nodes are ignored for now and merely regarded as hidden intermediaries for further word activation. In general, if activation is allowed to spread further – the activation spread setting is higher – more words are activated. Table 5 shows the average number of words that were activated by SIAM on reading the four passages containing each prime word. Small standard deviations show that the particular passage did not matter very much to the activation pattern, which is expected as all passages were normalized for length and position of the prime word.

However, as Table 5 also shows, high spread settings sometimes prevent any results from being returned. This is because so many reciprocal links or loops of links between nodes are created that when inhibition is applied in the final step, nodes are very quickly inhibited to below the minimum activation threshold, and are thus cut from the active node set. This is a shortcoming in the model, but it should also be noted the "spread setting" in the human lexicon would not be a one-time discrete choice as here,

²⁰ Of course, if the baseline activation of such nodes were low to begin with, this would be misleading. Such words might not have undergone inhibition from a negative inference, but by this evaluation standard might be considered as possible implicature items. However, as all nodes underwent inhibition in SIAM, we leave this question for a more detailed future version of the model. The words selected for evaluation here (typical scale words, as well as the other target words from the Chapter 3 experiments) were reasonably frequent and not subject to this concern.

but probably dependent on any number of factors, including experience, cognitive load, and so on. Certain levels of activation spread would be appropriate at different times. For this reason, results for a few different levels of activation spread are shown.

Prime	Spread=2		Spread=3		Spread=4	
	M	SD	M	SD	M	SD
BELIEVE	62	0.50	75	7.00		
COOL	22	0.58	24	5.00	46	19.09
KILLED	3	1.15	23	3.20	43	10.08
LIKES	3	0.00	20	0.00	47	4.32
PRETTY	7	0.00	22	1.00	44	4.65
SIMILAR	13	0.00	18	0.00		
SOME	69	0.00	143	9.81	118	18.41
SOMETIMES	16	5.85	60	3.59	123	29.01
STARTED	23	0.50	36	2.00	143	8.38
TWO	46	0.50				
WARM	18	2.00	37	7.55	59	

Table 5. Average number of words that are linked to the prime word, at different activation spread settings.

Possible scales

The primary question for this model is whether it would be possible for "scalar inference" to take place without a specific mental organization for a scale. To this end, it is necessary to assess which words might be the most likely targets for an implicature – meaning those that a person might be conscious of negating on hearing a lower scalar (or otherwise related) word. Here, we might look at either (i) words that have unexpectedly low activation given their frequency, or (ii) words which have the greatest change in activation from when they were first activated to the end of the passage's processing. The

results for either measure are the same, however, so we focus discussion on the former. In Chapter 3, entailment and word association were examined as factors for words negated via scalar implicature, and thus the target words with those manipulations are the focus here.

First, note that, with a very few exceptions, SIAM linked neither Non-Entailing/Non-Associate (NE/NA) words nor the Entailing/Non-Associate (E/NA) words to any of their respective prime words.²¹ That is, neither "chilly" (E/NA) nor "nice" (NE/NA) were activated by "cool" when SIAM read contexts with the prime "cool". It is possible that words of this type were activated and were subsequently inhibited below threshold. Entailing/Associate (E/A) words and Non-Entailing/Associate (NE/A) words, on the other hand, were almost always activated by their primes, at at least one activation spread setting. Table 6 shows how the activation of E/A and NE/A word nodes compare to the activation of the other word nodes that were activated for their respective primes. For instance, for the prime "some", 90% of the words that were activated and linked to "some" had greater activation than "all", averaged over the four different "some" contexts. Thus "all" was among the most inhibited words that were linked from "some".

Table 6 also shows the SIAM level of activation spread from where this average was drawn, as the target's activation relative to other words was variable across spread settings. As SIAM does not model spread settings in great detail, this table shows the percentage for the spread setting that had the lowest activation for the target word. Crucially though, the E/A and NE/A words were almost all linked to their respective primes and inhibited in at least one spread setting.

²¹ A full list primes and targets appears in the appendix.

Prime	E/A target	Activation ranking	Spread setting	NE/A target	Activation ranking	Spread setting
BELIEVE	know	0.00%	2	truth	54.28%	3
COOL	cold	73.91%	4	hot	88.80%	2
KILLED	murder	53.95%	4	destroy	28.29%	3
LIKES	love	75.41%	4	hate	52.88%	4
PRETTY	beautiful	77.78%	3	ugly		
SIMILAR	same	40.00%	2	different	75.00%	3
SOME	all	89.66%	3	none		
SOMETIMES	always	70.18%	3	maybe	89.46%	4
STARTED	finish	48.68%	2	stop	30.91%	3
TWO	three	65.57%	2	number	38.75%	2
WARM	hot	81.79%	2	cold	92.86%	3

Table 6. Ranking of E/A and NE/A target word nodes, as a percentage of all others linked to the same prime word, ordered by activation level. The percentage refers to the number of linked word nodes that have activation higher than the target word. A target ranking of 0% indicates that no other word node had higher activation, while a target ranking of 100% indicates that all the other linked word nodes had higher activation. In other words, a target with a ranking of 100% would have the lowest activation of all linked words.

E/A and NE/A word nodes were, on average, activated and inhibited at the same level. The average ranking for both is about 61%, indicating that 61% of the all the word nodes linked to the prime had greater activation than the target word node. By the proposed standards, then, these word nodes are among the ones that are most likely to have been negated via implicature.

Connections between nodes

Another way of looking at the data is to what degree words are interlinked. Table 7 shows the average count of nodes that directly activated the nodes of the target words, by

their entailing or associate category; in other words, how many parents each target node had. This includes parents besides the prime word, because any link to the target word provides an opportunity for a change in activation, either positively, by adding a new grandparent or great-grandparent node, or negatively through inhibition from some parent node. Table 7 thus reflects both the degree of word association / word frequency, as well as degree of semantic generality. Words with many parent links have a better chance of remaining active above the activation threshold.

Note that E/NA words have comparatively few links, reflecting both their lower frequency and semantic specificity. NE/NA words are as well linked as others, but as they have no semantic or associative link to the prime. Thus neither E/NA nor NE/NA words are likely to be perceived as targets of implicatures.

Prime	E/A words		E/NA words		NE/A words		NE/NA words	
	M	SD	M	SD	M	SD	M	SD
believe	8.00	6.26	1.00	0.00	5.33	1.15	12.75	8.96
cool	7.75	2.36			8.60	2.30		
kill	3.83	0.41			0.83	0.41	10.67	8.08
like	11.33	5.47			2.83	0.75	7.67	3.78
pretty	2.00	0.00					1.40	0.55
similar	9.00	1.41			7.50	1.29		
some	6.63	2.33	1.00	0.00				
start	5.33	5.82	1.33	0.58	10.17	5.95	3.17	2.99
two	4.00	1.15	1.67	0.52	13.00	1.41	1.00	0.00
warm	9.00	1.41			8.67	1.53		
All	6.38	4.43	1.25	0.44	6.40	4.53	6.04	6.19

Table 7. Number of nodes from which the target word nodes were linked (or, number of parents of each target word node) averaged across all spread settings and all texts.

Discussion

SIAM shows that many typical higher scalar words are especially sensitive to scalar inference because they have multiple word associations and semantic associations, and because the words' overall frequency makes those links strong. However, due to those same factors, it seems that even when such words are inhibited via negative scalar inference, they remain detectable, and are thus perceived to be negated via implicature. We suggest that there must be some detectability threshold below which this ceases to be the case, or some way of comparing previous activation to current activation. While this is quite speculative, it does seem reasonable to expect that some level of activation for a node must remain to perceive scalar implicature, as people are conscious of them, can recall making them, and so on.

To apply a specific example, a person might well have heard "cool" and inhibited "chilly" via a negative inference, as "chilly" is fairly infrequent and has few (but presumably not zero) links from "cool". However the activation of "chilly", a low-frequency word, would be reduced too sharply, or too far, for anyone to be conscious of thinking "not chilly". On the other hand, more active, interlinked nodes representing "cold" and "hot" remain, with low activation, where a negative inference would have had the most implicature-like effect.

Of course, the results showed that the target words were on average in the bottom 40% of active nodes that were linked to the prime. Many other words, besides the target words evaluated here, were similarly inhibited. Why is scalar implicature not available for these or any merely associated words? For instance, "maybe" was among the most

inhibited words from the prime "sometimes". Why is there no possible implicature from a scale like <maybe, sometimes>? ("Jane sometimes loses her keys. In fact, she maybe does.") The overall context, and part of speech of the word as used in the sentence, must make a difference. Again, scalar implicature is not a purely lexical phenomenon. World knowledge about a given proposition is what triggers inferences, and presumably the interaction between the sentence/proposition and lexical activation is relevant for what implicatures people are aware of. However, given a pair of any words (propositions) with sufficiently strong association and activation, and something like scalar implicature would be possible.

SIAM also does not distinguish between Entailing/Associate targets and Non-Entailing/Associate targets (again, for prime "cool", these are "cold" and "hot" respectively), and yet NE/A words are not part of any typical generalized scale, whereas E/A words are. What seems likely is that the impression of a scale results from processing the rest of some context with implicature, and from ongoing reasoning. For instance, given the implicature-supportive context "You don't need anything heavy to wear. It's cool outside," hearers' attention is called to "cold" in part because they know that people wear heavy clothing in the cold. "Cold" does also logically entail "cool". We know from scales like <succeed, try>, this is not a firm requirement for something to be called scalar implicature, but it leads to support for the idea of the mental representation of a scale.

It is easy to think of a comparable context for "hot": "The summer has been unbearable, but it's cool today." This is more likely characterized as contrast than implicature. Still, results from SIAM suggest that the same lexicon organization is

behind the inference that the speaker means to convey that it is not hot today. Of course, this is facilitated by other knowledge, e.g. a day can't be both cool and hot. But nobody would claim that <hot, cool> is a scale in the sense of scalar implicature.

The model overall seems to suggest that, basically, word frequency, association to a particular word, the presence of additional semantic links to that and other words, world knowledge, and the context of a sentence can account for the seemingly scalar nature of scalar implicature. A pre-existing mental structure for defined lexical scales is not necessary.

Conclusions

SIAM has many limitations, but does seem plausible as a model for the lexical aspect of scalar implicature, that is, what words are thought of as being negated via scalar inference. Although SIAM returns the widest possible set of such words, which must be winnowed down through knowledge-based inferences and syntactic input, what is crucial is that there is no need for a generalized implicature heuristic, or a mental construct that links a specific set of words by some ordering principle.

It is true that theoretical discussions of scalar implicature have generally not made specific claims about underlying lexicon structure. It is understood that scales are an abstraction. Researchers in pragmatics have not necessarily committed to the idea of an actual mental structure for them. The usefulness of SIAM is in showing explicitly that such a mental structure is not necessary. If that is the case, then scalar implicature should

generalizable past logically entailing words, as is often reported but not often accounted for.

SIAM further explains why scalar implicature is not possible for any one word that entails another. An argument for why some scales exist, and some do not, is that words must be about the same concept, have more or less the same number and type of semantic features (Levinson 2000). In some sense, SIAM is an implementation for that idea. However, it further reveals that words that are negated via implicature must not merely be of the same order of semantic specificity, but must always be more general. A scale <freezing, chilly> does not exist (without some particular context), even if those words would be considered on the same semantic level. SIAM does not link such words because they are relatively infrequent, and they are semantically separated by a relatively large distance.

The primary hope for this model is to re-motivate efforts to unify different types of scalar implicature. For instance, implicatures regarding terms that are not ordered, like colors, days of the week, or conjunctive phrases, are often thought of as scalar though they may have no regular ordering. In addition, particularized scalar implicatures that require a great deal of world knowledge might function as in SIAM, where nodes representing whole propositions ("I like Greek food"; "I speak Greek") are activated and then inhibited. It might be well to reduce the emphasis on scale in scalar implicature.

Chapter 5: Conclusions and Further Research

The goal of this dissertation has been to address existing questions of scalar implicature processing, and to introduce new ones. I will begin this concluding chapter by summarizing my contributions to the former, with some suggestions as to their implications for pragmatic theory.

Cost and automaticity

The experiments of chapters 2 and 3 provide evidence in favor of effortful processing associated with scalar implicature. Both demonstrate an extended response time in implicature-supportive contexts. In chapter 3, there was a strong main effect of context, such participants were overall slower to respond to target items after hearing implicature-supportive contexts. In the eyetracking experiment of chapter 2, both participant response times and fixation times (periods of time during which the eye was gazing at a particular point in a scene) were longer when participant heard the word "some", and was therefore aware that an inference "not all" was possible and perhaps intended. Longer fixation times are generally considered indicators of greater cognitive load (Inhoff and Radach 1998). This dissertation thus affirms much previous research in suggesting that drawing an inference from scalar implicature requires time-consuming, active cognition.

An assumption underlying many studies, particularly those claiming evidence for Relevance Theory, has been that if such cognition is to be undertaken, it must be due to some choice or evaluation by the comprehender. This has been the source of the strong

dichotomy of Default/free and Relevance/costly theories of implicature processing. In a way this dichotomy has been false to begin with. For instance, people automatically process the language that they hear, which always requires at least some effort and also carries a possibility that a sentence may turn out to be complex or difficult to understand. It is the default to attempt to understand linguistic input. The (rarely-made) choice may be to abandon the processing if the effort gets to be too much, but by default, comprehenders extend effort towards comprehension. There is no reason then to assume that implicature processing must be free in order to be automatically undertaken.

The idea of any inference processing being free is a simplification that is likely not anyone's true intention. "Free" has always meant "so cheap as to be essentially free", not "completely without cost of any kind." When placed in these terms, the question of automaticity revolves around a difference of thresholds – exactly how cheap is that? Where is the default threshold of effort past which a hearer will not process an implicature? While the Default view states that there is little effort associated with implicature processing, having demonstrated effort, one might revise the view to suggest that the default threshold is simply extremely high, and thus implicatures will always, automatically, be processed.

The problem with this revision is that it would propose a different threshold for different types of inferences. For Generalized Conversational Implicatures (GCIs) having particular well-defined scales, the threshold is high, and these are thus invariably processed, by Default. For inferences from other sources, the threshold is lower, perhaps variable. Leaving aside troubling questions regarding the implementation of this privileged class of inferences, the new Default view is now problematic in that it is little

different from what Relevance Theory proposes: different processing thresholds for different circumstances.

What is clear is that inferences from scalar terms are not automatically made, but depend (at least) on context. While the chapter 2 experiment suggests that the "not all" inference was always made when participants heard "some", the chapter 3 experiments, along with previous experiments, demonstrate statistically significant differences in response times between implicature-supportive and non-implicature-supportive contexts.²² If it is possible to draw a line between the behaviors elicited by these two types of stimuli, it seems unlikely that one can claim that any class of inferences is inherently automatic. The stimuli of chapter 2 would simply fall into the implicature-supportive category.

The computational model of chapter 4 has proposed that words forming typical GCI scales are likely to be strongly related (interlinked, associated) in the lexicon. To a degree this also implies automaticity, in that higher scalar words are automatically primed and activated when the corresponding lower scale words are heard. To this extent, implicature processing may be automatic. But this extent is not very far. The negation of these higher-scale items, manifested as inhibition, reflects a secondary step that is made after some very complex determination involving world knowledge, immediate context, hearer fatigue, and so on. This calculation, depending on so many factors, cannot meaningfully be said to be automatic, or to have any default setting.

²² The other suggestion of Levinson (2000) is that implicatures are processed by default, but then are cancelled if they are not appropriate to the context. In my view, the simpler explanation is that they simply aren't generated. Furthermore, there are no predictions about the nature of cancellation processing .

Global vs. local processing

The debate over whether implicature processing occurs locally at the scalar term (or in psycholinguistic terms, incrementally as the parsing proceeds), or globally after an entire utterance, has been of primary interest to many theorists. This is not as true for psycholinguists, who tend to be unsupportive of the globalist view with its strongly serialized theories of processing. Thus the present work, for the most part, was not designed with this theoretical debate in mind. More to the point, it did not use stimuli with distinct globalist/localist predictions, such as multiple scalar items, embedded scalar items (especially under factives), or “multiplicatures” (Geurts 2007). However, while the current research is not definitive in distinguishing between the two views or between differing localist accounts, the results do support psycholinguistic interactionist views, in which all manner of factors, but mainly contextual, guide the parse. The idea is briefly outlined in Breheny, Katsos and Williams (2006), and has something to do with both global and local views, but is hard to completely describe within pragmatics. Since the theoretical accounts of global and local views are incompatible, though, I must overall conclude that the localist view is more correct. I found no support for the strict version of global implicature computation where pragmatic or contextual information has no effect on the initial parse, and a completely worked out interpretation is secondarily passed to a pragmatic module.

Still, addressing psycholinguistic research from the perspective of this debate can yield insights. For instance, the global/local distinction might have had significance for the eyetracking experiments reviewed in chapter 2, specifically Grodner et al. (2010) and

Huang and Snedeker (2009). Both studies had stimuli along the lines of “Point to the girl with some of the balls” (or “socks” for H&S), which, if it were interpreted as an existential statement, would have one of the following implicatures:

(14) Global calculation: Point to the scene where it is not the case that there is a girl with all of the (balls / socks).

Local calculation: Point to the scene where there is a girl with not all of the (balls / socks).

Having a phonetic competitor (balloons / soccer balls) complicates matters, but it is interesting to note that the globally-computed implicature would not have uniquely identified a target. There were two mini-scenes where it was not the case that the girl had all the X: the girl with the 2 balls, and the girl with nothing. However, under a locally-computed version, there was only 1 target: the girl that had 2 balls (at least, assuming that “There is a girl with not all the X” could not be interpreted as the girl who has nothing at all, which seems likely.) The Huang and Snedeker scene is similar. There are two scenes where it seems true that it is not the case that the girl has all the socks (girl with 2 socks, girl with soccer balls), but only one scene has a girl with not all of the socks.

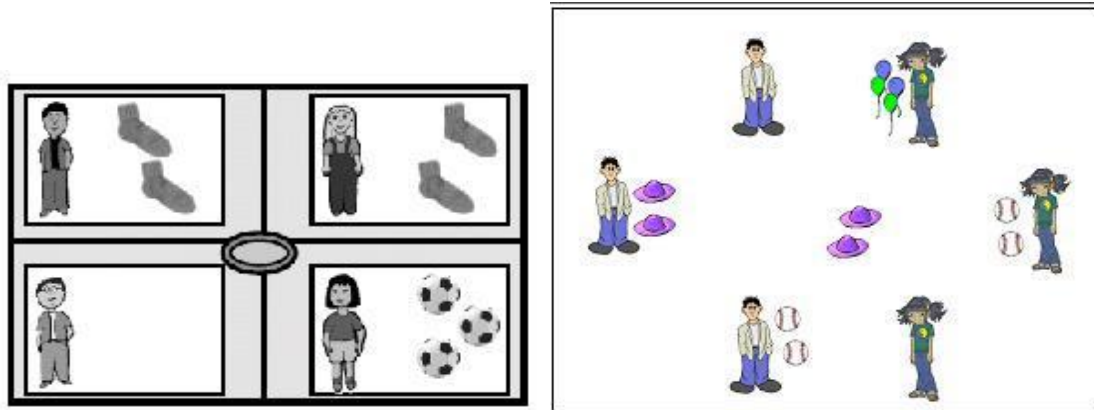


Figure 14. Sample visual scenes from Huang and Snedeker (2009) (left) and Grodner et al. (2010).

This might be taken as evidence for local computation, since under the global view, the comprehender might be equally likely to look at either of the two scenes. An eyetracking experiment specifically comparing the global and local computation of implicatures might be designed in this way.

The chapter 2 experiment can only contribute indirectly to this debate.²³ The stimuli, of the form “Some of the apples are red”, had the scalar term in the subject position. This would have the same predictions under either a global or local account of implicature. Again, under the global view, alternative propositions are generated only after the grammatical structure of the sentence has been completely computed.

Alternatively in the local view, where strengthening of the sentence takes place from the bottom up, the addition of scalar alternatives would not take place until “some” was

²³ The computational model of chapter 4 also contributes slightly in that it assumed local implicature processing by generating/negating alternatives occurred immediately after the activation of the original word. Again, though, the goal of the model was to demonstrate why certain lexical items could form scales, and not others. With this lexical focus, and the overall lack of structural information, it would not have made very much difference if the processing had occurred at the end of the sentence. However, future modifications to the model could incorporate both syntactic parsing and finer-grained parameters, with the goal of demonstrating the importance of propositions in implicature processing.

finally integrated at the top of the structure. In both cases, alternatives to “some” would not be created until late.

Interestingly, the results that address the issue at all are at odds with those predictions. When participants heard “some”, and there were two groups of differently colored objects on the screen, they looked at the “other” group, (relative to the one they had been looking at) quickly – before the sentence was completely parsed. This did not happen as quickly when the quantifier was “all” with the same picture, and the looks were not triggered by the color word, because they preceded it. The conclusion was that participants immediately inferred “not all”, and since they (probably) knew from previous experience that the items would be distinguished by color and that the only difference in the items was their color, they looked at the evidence that supported the inference “not all of them are the same color”. The likely explanation is that the scalar alternatives are generated after the some preliminary, temporary parse, not a complete bottom-up parsing of the completed sentence. If we allow for this possibility, there is no longer any way to distinguish between the formal global and local views.

Nature of scales

The first chapter introduced some of the concerns about what may constitute a scale in scalar implicature. The most typical scale discussed in pragmatic theory, the Horn scale, consists of semantically general terms, where the higher terms logically entail the lower ones. Most psycholinguistic experiments have limited the definition further by using scales where there is only a logical relationship, as with numbers or quantifiers, as

opposed to scales with an additional ordering parameter besides entailment, such as degrees of temperature (<cold, cool>) or intention (<murder, kill>). On the other side of the issue, theorists have suggested there cannot be a scale of lexical items that are of greater (or variable) semantic specificity, and also that pseudo-scales of specific propositions or predicates (<speaking Greek, liking Greek food>) are of a different category altogether, namely, particularized implicatures.

The open question is whether, and to what degree, these different scale types should be treated similarly. In my view, the only possibly legitimate distinction is between scales of closed-class words and scales of open-class words, which was perhaps the spirit behind Generalized Conversational Implicatures in the first place. There seems to always be some additional complication of ordering for open-class words, e.g., the degree to which a word is semantically specific. That said, however, neither experimental data of chapter 3 nor the computational modeling of chapter 4 suggests a need for an open-class/closed-class distinction in scales. In fact, there is possibly no need to distinguish between any set of words or propositions that are orderable in a scalar way.

The experiments of chapter 3 showed that, in contexts that supported scalar implicature, there were apparent effects of implicature processing not only on words that entailed the lower scalar item in the stimuli, but also on those words that were associated with it. In addition, chapter 4 showed that any words that had a sufficient number of sufficiently strong connections to a "lower" word could be negated via a scalar inference. This explained why frequent, semantically-general words make up the typical GCI scales, and why scalar implicature does not seem possible with more specific words. The point I would like to emphasize, though, is that the same type of structure and procedure

modeled in SIAM could be used to explain why strongly context-based "particularized" scales exist, and why they indeed have something in common with lexical scales.

I suggested in chapter 4 that the perception of scales is the effect of post-hoc reasoning and analysis. It is certainly true that people are very quick to perceive that something is "more" in some sense than something else, or that some word represents more of something than another word. My belief is that this tendency, applied so often to high-frequency words, has led to the idea of scalar implicature. However, this is not to say that scalar implicature is not a real phenomenon worth investigating, only that the formalization of scale composition is probably pointless. My hope is that these findings will motivate research into a now very generalized type of scalar implicature, with more effort ultimately devoted to discovering the underlying mechanics for all scale types.

Methodological issues

Besides the nature of the task (discussed in chapter 2), there are other issues that should be considered in future experimentation on scalar implicature, and subsequent data analysis. It appears that the effectiveness of experimental trials using implicature-supportive stimuli varies when comprehenders are exposed to them repeatedly. In chapter 2, I found an apparently strong initial difference between two conditions, one which was intended to cause implicature processing and one which could have but was not expected to. However, an examination of the individual trials comparing these conditions showed that the difference was much weaker after the initial trial of the conditions. Similarly, in chapter 3, the results were only significant when examining the

first trials for each condition, and in some cases the second trials. By the third and final trials, the effects had disappeared.

Thus when results are averaged over repeated trials in a condition (for a given participant), they may be misleading. This may be due to a number of different factors. In chapter 2, I speculated that participants were initially boggled by the task of determining whether a description was good or not (e.g., good in whose opinion?), and that as this decision did not need to be repeated over five trials, the response times became faster. In chapter 3, the short-lived effects were likely due to some "priming" of the scalar inference. Participants frequently noticed the relationship between target and prime, and while almost nobody was conscious of having processed an implicature, the data suggests that the scalar inference was made. Having been made repeatedly, it would become faster.

Because of the statistical significance of the chapter 2 and 3 results, I do not believe these issues invalidate the experiments. Rather I suggest that future experiments on scalar implicature should attempt to use a larger number of participants, with fewer repeated trials, in order to eliminate the instances of participants' use of decision making strategies (e.g. deciding what the experimenter thinks is a good or true answer) or priming/learning effects from repeated inferences. Alternatively, experimenters should be careful to look at the results from individual trials for any condition with repeated trials. While implicature processing is costly, it appears to become cheaper quickly.

Thoughts for future research

There remain a number of questions related to scalar implicature to which I have not given due attention. Among them is the absence of scalar implicature, or reversal of scales, under negation. (Atlas and Levinson 1981; Horn 1989):

- (15) a. Some of the apples are red.
b. It is not the case that some of the apples are red.
c. It is not the case that all of the apples are red.

Sentence (15b) seems to have no implicature. A speaker uttering it would not seem to necessarily mean that either all or none of the apples are red, or anything else related to sentence (15a). Sentence (15c) on the other hand, where a higher scalar item is negated, does seem to at least potentially imply that some of the apples are red. The scale has been reversed; the negation higher item "all" implicates "some".

My tentative thought is that examples like sentence (15b) could be explained by inhibition. In chapter 3 I discussed evidence suggesting that negation of a word might cause it to be inhibited, and similarly, that the negative inference drawn on processing a scalar implicature might lead to the inhibition of the words in the inference. In other words, hearing "some" leads to the inference "not all", which causes "all" to be inhibited. If there were further negation above "some", however, the activation that would propagate to "all" would be reduced even further, and thus the inference would effectively not have happened. This is a very speculative explanation, and it is not

helped by the apparent reversal of scales in sentence (15c); inhibition of "all" should not then cause "some" to be more active, as it apparently does.

On the other hand, it is not clear how robust the scale-reversal effects are. If a person says out of the blue "It's not cold outside", a hearer would not immediately conclude that it was either cool or hot, but only that there must be some additional intended meaning (e.g. that the speaker thought that the hearer thought it was indeed cold.) That meaning might be provided by a greater context: "You might want a coat, though it's not cold outside." In that case, though, it is hard to say what exactly is responsible for the implied meaning. The only point I wish to make for now is that the evidence from negation would likely be a fruitful direction for research

The most basic question is, why should scalar implicature happen at all? Levinson (2000) described it as a heuristic such that speakers may convey meaning without going to the trouble of saying it. Yet this view merely transfers the burden (nonexistent in Levinson's view, but demonstrated since) to the hearer. It is also unreliable as a heuristic; even the most generalized scalar implicature depends on context to some degree. Finally we have seen that words that often make up scales are associated in the lexicon, which suggests that there are common enough circumstances where the words must be used explicitly. In my view, scalar implicature does not present any strong benefits for language processing.

Relevance Theory maintains a general-cognition view, in that all language processing has the same motivation: to derive as much information as possible from input, given the effort available. This is intuitively attractive in its characterization of humans as information-seeking beings, but ultimately is no more satisfactory. Its main

problem is that it presumes that the hearer is able to quantify both available effort and available information, without explaining how either is possible. More importantly, though, it is questionable whether generalization ought to be the ultimate goal. The brain is specialized and localized for language, and even for different aspects of language, as well as its many other functions. To describe all language processing as information-seeking might be basically true, but seems reductive past the point of usefulness.

Of course, there may be no real need to find a reason behind scalar implicature, but insofar as it is interesting to speculate, I would like to maintain a linguistic explanation of a linguistic phenomenon. Relevance Theory proposed that the Maxim of Relevance should be the basis for all language processing, but my thought is that perhaps the Quantity maxim is more explanatory. If we can accept the idea that scalar implicature processing causes the active suppression of linked words (or propositions), both stronger and not, then in the broader sense, we are saying that hearers believe that speakers mean exactly what they say, and do not mean what they didn't say. This is suggestive of a processing benefit also. When fewer alternative words are active, semantic and pragmatic processing is potentially faster. Initially, when a person hears a word, many alternative related words become more active. Over time, though, it may not be that activation of these words merely fades, but rather that it is actively suppressed by some Quantity-based motivation to limit the amount of data that must be considered in complex, contextually-based language comprehension.

Bibliography

- Altmann, G. T. M. & Steedman, M. J. (1988). Interaction with context during human sentence processing. *Cognition*, 30(3), 191-238.
- Atlas, J. & Levinson, S.C. (1981). It-clefts, informativeness, and logical form. In P. Cole (ed.), *Radical Pragmatics* (pp. 1-61). New York: Academic Press.
- Bezuidenhout, A. L. & Morris, R. K. (2004) Implicature, relevance, and default pragmatic inference. In I. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp 257-282). Basingstoke, Hampshire, UK: Palgrave Macmillan.
- Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language*, 37, 83-96.
- Bott, L. & Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51, 433-456.
- Breheny, R., Katsos, N., & Williams, J.(2006). Are generalized scalar implicatures generated by defaults? An on line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100, 1-30.
- Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977-990.
- Cassidy, K. W., Kelly, M. H., & Sharoni, L. J. (1999). Inferring gender from name phonology. *Journal of Experimental Psychology*, 128, 362-81.
- Clifton, C., Jr., Staub, A., & Rayner, K. (2007). Eye movements inreading words and sentences. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341-371). Amsterdam: Elsevier, North-Holland.
- Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (Ed.), *The cartography of syntactic structures*. Vol. 3, *Structures and beyond* (pp. 39-103). Oxford: Oxford University Press.
- Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the "logicality" of language. *Linguistic Inquiry*, 37(4), 535-590.
- Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6, 84-107.

- Davies, M. (2008). Corpus of Contemporary American English [Database]. Retrieved from <http://www.americancorpus.org>.
- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, *54*, 128-133.
- Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C. & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*, 409-436.
- Fellbaum, C. (Ed.). (1998). *WordNet. An Electronic Lexical Database*. Cambridge: MIT Press.
- Foppolo, F. (2007). Between "cost" and "default": a new approach to scalar implicature. In R. Artstein, & L. Vieu (Eds.), *Decalog 2007: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 125–131).
- Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland & P. Stateva (Eds.), *Presupposition and implicature in compositional semantics* (pp. 537–586). New York: Palgrave Macmillan.
- Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.
- Geurts, B. (2009). Scalar implicature and local pragmatics. *Mind and language*, *24*, 51-79.
- Grice, P. (1975). Logic and conversation. In Cole, P. & Morgan, J. (Eds.) *Syntax and semantics 3: Speech acts*. New York: Academic Press.
- Grodner, D., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). “Some,” and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, *116*(1), 42-55.
- Henderson, J. M. & Ferreira, F. (2004) Scene perception for psycholinguists. In Henderson, J. M. & Ferreira, F. (Eds.), *The interface of language, vision, and action : eye movements and the visual world* (pp. 1 – 58). New York: Psychology Press.
- Hirschberg, J. (1985). *A theory of scalar implicature*. Doctoral dissertation, University of Pennsylvania.
- Horn, L. (1972). *On the semantic properties of logical operators in English*. Doctoral dissertation, University of California Los Angeles.

- Horn, L. R. (1989). *A natural history of negation*. Chicago: University of Chicago Press.
- Huang, Y. & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, 58(3), 376-415.
- Inhoff, A. W. & Radach, R. (1998) Definition and computation of oculomotor measures in the study of cognitive processes. In Underwood, G. (Ed.), *Eye guidance in reading and scene perception*, (pp. 29-75). Oxford: Elsevier Science.
- Katsos, N., Breheny, R. & Williams, J. (2008). Interaction of structural and contextual constraints during the on-line generation of scalar inferences. In *Proceedings of GLOW 28*. University of Geneva.
- Kaup, B. (2001). Negation and its impact on the accessibility of text information. *Memory and Cognition*, 29, 960-967.
- Levinson, S.C. (1989). A review of Relevance. *Journal of Linguistics*, 25, 455-472.
- Levinson, S. C. (2008). *Presumptive meanings*. Cambridge, MA: MIT Press.
- MacDonald, M. C., & Just, M. A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 15, 633-642.
- MacDonald, M. C. & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd ed.) (pp. 581-612). London: Academic Press.
- Matin, E., Shao, K. C., Boff, K. R. (1993) Saccadic overhead: information processing time with and without saccades. *Perception and Psychophysics*, 53(4), 372-380.
- McNamara, T. P. (2005) *Semantic priming: Perspectives from memory and word recognition*. New York: Psychology Press.
- Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1), 91-117.
- Mirković, J., MacDonald, M. C., & Seidenberg, M. S. (2005). Where does gender come from? Evidence from a complex inflectional system. *Language and Cognitive Processes*, 20, 139-168.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from <http://www.usf.edu/FreeAssociation/>.

- Noveck, I. (2001). When children are more logical than adults: Experimental investigations of scalar implicatures. *Cognition*, 78, 165–188.
- Noveck, I. & Posada, A. (2003) Characterizing the time course of an implicature. *Brain and Language*, 85, 203-210.
- Papafragou, A. & Musolino, J. (2003) Scalar implicatures at the semantic-pragmatics interface. *Cognition*, 80, 253–282.
- Perea, M., & Gotor, A. (1997). Associative and semantic priming effects occur at very short SOAs in lexical decision and naming. *Cognition*, 62, 223-240.
- Postman, L., & Keppel, G. (Eds.) (1970). *Norms of word association*. New York : Academic Press.
- Pylkkänen, L., & McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In M. Traxler & M. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd ed.) (pp. 539-580). London: Academic Press.
- Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology*, 55A, 1339-1362.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124 (3), 372
- Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1, 75-116.
- Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of Semantics*, 23, 361–382.
- Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G. & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109–147.
- Storto, G. & Tanenhaus, M. K. (2005) Are scalar implicatures computed online?. In E. Maier, C. Bary, & J. Huitink (Eds.), *Proceedings of Sinn und Bedeutung* (pp. 431-445).
- Storto, G. & Tanenhaus, M.K. (2007). Are scalar implicatures computed online? In A. Alcazar, R. Mayoral Hernandez & M.T. Martinez (Eds.), *Proceedings of the Western Conference on Linguistics*. Fresno: California State University at Fresno.
- Spell, B. (2008) Java API for WordNet Searching [Computer software]. Retrieved from <http://lyle.smu.edu/~tspell/jaws/index.html>

Sperber, D. & Wilson, D. (1995) *Relevance: Communication and cognition* (2nd ed.). Oxford: Blackwell Press.

Steedman, M. J. & Altmann, G. T. M. (1989). Ambiguity in context: A reply. *Language and Cognitive Processes*, 4(314), 105-122.

Tanenhaus, M. K. and Trueswell, J. C. (2004). Eye movements as a tool for bridging the language-as-product and language-as-action traditions. In M. K. Tanenhaus & J. C. Trueswell (Eds.), *Approaches to studying world-situated language use* (pp. 3-38). Cambridge: MIT Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632-1634.

Tanenhaus, M. K. & Trueswell, J. C. (2006). Eye movements and spoken language comprehension. In M. Traxler & M. Gernsbacher (Eds). *Handbook of psycholinguistics* (2nd ed.), (pp. 863-900). Amsterdam: Academic Press.

Toutanova, K. & Manning, C.D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, (pp. 63-70).

Toutanova, K., Klein, D., Manning, C. D. & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

Appendix A

Chapter 2 stimuli

Table 8. Target items.

Item	Audio stimuli
1	This is a picture of apples. (Some/all) of them are red.
2	This is a picture of bags. (Some/all) of them are brown.
3	This is a picture of birds. (Some/all) of them are brown.
4	This is a picture of boats. (Some/all) of them are orange.
5	This is a picture of books. (Some/all) of them are yellow.
6	This is a picture of bottles. (Some/all) of them are green.
7	This is a picture of cars. (Some/all) of them are pink.
8	This is a picture of cats. (Some/all) of them are gray.
9	This is a picture of cups. (Some/all) of them are blue.
10	This is a picture of dogs. (Some/all) of them are black.
11	This is a picture of eggs. (Some/all) of them are white.
12	This is a picture of flowers. (Some/all) of them are yellow.
13	This is a picture of hats. (Some/all) of them are blue.
14	This is a picture of horses. (Some/all) of them are black.
15	This is a picture of houses. (Some/all) of them are pink.
16	This is a picture of shoes. (Some/all) of them are red.
17	This is a picture of tables. (Some/all) of them are gray.
18	This is a picture of candles. (Some/all) of them are purple.
19	This is a picture of coats. (Some/all) of them are green.
20	This is a picture of shirts. (Some/all) of them are white.

Table 9. Filler items.

Item	Audio stimuli
21	This is a picture of acorns. There are 4 brown ones.
22	This is a picture of ashtrays. The blue ones are on the right.
23	This scene shows bears. They are angry.
24	This is a picture of beds. There are 5 indigo ones.
25	This is a picture of bells. They are silver.
26	This scene shows bikes. There are 4 cobalt ones.

Table 9. (Continued)

Item	Audio stimuli
27	This is a picture of blinds. They are open.
28	This is a picture of boxes. The brown ones are on the bottom.
29	This scene shows circles. There are 2 navy ones.
30	This is a picture of clouds. The white ones are on the bottom.
31	This is a picture of compasses. They are gold.
32	This scene shows dishes. The orange ones are on the top.
33	These are models of the Eiffel tower. They are vermilion.
34	This is a picture of gates. They are closed.
35	This scene shows hanggliders. There are 4 violet ones.
36	This is a picture of hearts. They are ivory.
37	This is a picture of ladybugs. They are smiling.
38	This scene shows leaves. There are 4 red-brown ones.
39	This is a picture of lightbulbs. They are light yellow.
40	This is a picture of lions. They are happy.
41	This scene shows nails. They are silver.
42	This is a picture of sandwiches. They look fresh.
43	This is a picture of octagons. There are 2 orange ones.
44	This scene shows padlocks. They are violet blue.
45	This is a picture of pills. There are 2 light blue ones.
46	This is a picture of pots. They are aluminum.
47	This scene shows rakes. They are gold.
48	This is a picture of rockets. They are blue.
49	This is a picture of screwdrivers. The purple ones are on the left.
50	This scene shows seashells. They are beige.
51	This is a picture of stars. There are 4 shiny ones.
52	This is a picture of statues. There are 8 gray ones.
53	This scene shows tickets. The dark pink ones are on the right.
54	This is a picture of ties. There are 4 maroon ones.
55	This is a picture of toasters. They are black.
56	This scene shows wheels. There are 3 white ones.
57	This painting is the Mona Lisa.
58	This painting is Washington crossing the Delaware.
59	This flag is the Spanish flag.
60	This flag is the Finnish flag.

Appendix B

Chapter 3 stimuli

Table 10. Target words used in Experiments 1-3 in Chapter 3. Data elicited by shaded word "bump off" was excluded. Capitalized words are a different part of speech than the prime.

Prime	Entailing Associate (E/A)	Entailing Non-Associate (E/NA)	Non-Entailing Associate (NE/A)	Non-Entailing Non-Associate (NE/NA)
believe	know	swear	TRUTH	think
cool	cold	chilly	hot	nice
killed	murder	bump off	destroy	leave
likes	love	adore	hate	want
or	and	with	EITHER	after
pretty	beautiful	gorgeous	ugly	ready
similar	same	identical	different	remarkable
some	all	bunch	NONE	both
sometimes	always	constantly	MAYBE	certainly
started	finish	complete	stop	like
two	three	six	NUMBER	zero
warm	hot	overheated	cold	pleasant

Table 11. Nonwords (see table below for associated filler items). (Rastle, Harrington & Coltheart 2002)

briccane	ledes	sause
ceast	momb	shier
dersed	muise	sligh
ghap	nec	thwaught
gheshed	pe	varf
ghuisery	phawner	wheeperg
gnip	phigh	whelte
goniger	phign	wocks
joar	renningly	wrisc
kares	rhane	yarreed
kede	rhodes	
knape	rhuill	
knix	sa	

Table 12. Basic contexts (non-implicature-supportive) used in Experiments 1 and 2 in Chapter 3.

Item ID	Prime	Text
1	believe	Certainly, I'll go upstairs and get Beth. I believe she's home.
2	cool	Bring something heavy to wear, because it's cool outside.
3	killed	The suspect killed Officer Smith.
4	likes	Andrea is a big fan of all kinds of sweets, and she's always got some with her. She likes chocolate.
5	or	Bob is trying to gain weight so he planned to eat a thick soup or a big sandwich.
6	pretty	They got a great deal on that house! It has four bedrooms, and it's pretty.
7	similar	The doctor isn't sure what's wrong with me. For all kinds of illnesses, the first symptoms are similar.
8	some	Ellen has been looking for her documents since this morning, though I haven't talked to her about it. By now she's probably found some of them.
9	sometimes	Our co-workers think Barbara has no friends to go out with. I've seen her at Starbucks by herself, so she is out alone sometimes.
10	started	At the beginning of the school day, the students started editing their essays.
11	two	My cat is going to have kittens. I hope she has two.
12	warm	Have some tea! My mother just made it, and it's probably warm.
13	believe	After studying years of data, I believe the air is getting cleaner.
14	cool	It's safe to eat the egg salad. The room has been cool.
15	killed	The gang member killed the witness before he could testify.
16	likes	Kyle's wife introduced him to Belgian beers, and now he's drinking them all the time. He definitely likes them.
17	or	Maria just got a raise, and now she has piles of money coming in. She'll definitely spend it. She'll buy a new TV or new furniture.
18	pretty	Bridget could be on a magazine cover. She is so pretty.

Table 12. (Continued)

Item ID	Prime	Text
19	similar	I can't tell those teen actresses apart. They all look similar.
20	some	I can't get the kids to clean up after themselves. I'll have to pick up some of the toys.
21	sometimes	In an argument with your spouse, it is sometimes better to think before you speak.
22	started	In the afternoon, the athletes started the race.
23	two	My friend has been fighting sickness all year. She really believes vitamins are the key to recovery. She'll take two vitamins today.
24	warm	The child looks terrible, like she has the flu. Her forehead is warm.

Table 13. Implicature contexts used in Experiments 1 and 2 in Chapter 3.

Item ID	Prime	Text
1	believe	I haven't seen Jane, and I don't see her coat or bag where it usually is. But I believe she's home.
2	cool	You won't need anything heavy to wear. It's cool outside.
3	killed	In the line of duty, Officer Smith killed the suspect.
4	likes	Carol doesn't have a big sweet tooth. She likes candy.
5	or	Bill is trying to lose weight so he eats as little as possible. For lunch he has a chicken sandwich or a turkey sandwich.
6	pretty	They paid way too much for that house. It's nothing special. It's pretty.
7	similar	The common cold and influenza are similar illnesses.
8	some	Sarah has been looking for her books this morning. She found some of them.
9	sometimes	Many people think Jane is not reliable. You can count on her sometimes.
10	started	In the last few minutes of the school day, the students started editing their essays.
11	two	Business is awful. By the end of the day we'll have two orders.

Table 13. (Continued)

Item ID	Prime	Text
12	warm	I wouldn't drink that tea. I don't know when it was made, and it's probably warm.
13	believe	I believe the repairs will be done by tomorrow.
14	cool	I think the air conditioner isn't working right. The air that it's blowing out is cool.
15	killed	The worker at the top of the telephone pole dropped one of his tools. He killed the man on the sidewalk.
16	likes	Joe's wife is a huge fan of Belgian beers. Joe likes them.
17	or	In Pennsylvania it rains and snows all year. Matt is moving there but all he has is a motorcycle, so because of the weather he'll need a car or a truck.
18	pretty	There's not much to say about Jen. She's an ordinary person. She's pretty.
19	similar	I thought that white tennis ball was a baseball. It was an easy mistake to make, because they look similar.
20	some	I don't see why I should pick up after a grown man. I'll pick up some of the dirty socks.
21	sometimes	John stays up late, and gets up early sometimes.
22	started	To show the students how to paint the wall, the teacher started with them.
23	two	My friend is on a strict weight-loss diet. Aside from her meals, she'll eat two snacks today.
24	warm	The child is perfectly healthy. Her eyes look fine and her forehead is warm.

Table 14. Additional text for lengthened contexts in Chapter 3 Experiment 3. The text here was prepended to both the Basic and Implicature contexts with the corresponding item ID.

Item ID	Additional text
1	It's about seven-thirty now and dinner will be served at eight.
2	According to the weather channel report, the weather has been clear everywhere.
3	The police were called to the downtown warehouse district because of a shooting.
4	There's a lot left over from the party, so we should give it to someone.
5	Pretty much everyone is concerned about their health right after the holidays.

Table 14. (Continued)

Item ID	Additional text
6	My sister and brother-in-law just sold their condo and bought in a subdivision in New Hampshire.
7	I need to take care of myself when you're sick, though it depends on what I have.
8	Sarah and Ellen are in the grad student dorm. Their room is clean and organized.
9	Barbara and Jane both lived in my building last year, below my apartment.
10	The principal of the middle school said that writing was very important.
11	It's about time for things to change. I can't wait for spring.
12	There are drinks on the table, along with milk and sugar.
13	It's been an awful lot of work and taken a long time.
14	This week has been extremely humid, especially today when it was almost 100%.
15	A man was on his way to court to make a statement about a crime he had seen.
16	European beers are made differently from American ones, but both are available here.
17	Matt and Maria are excited about living in new parts of the country.
18	Jen and Bridget both work in the accounting department of a big company.
19	Now that I'm getting older, I'm sure that my vision is getting worse.
20	The house gets to be such a mess, especially on the weekends.
21	It's hard to predict how people are going to behave, even if you know them.
22	Everyone did a lot of practice ahead of time to make sure they were ready.
23	A lot of people are getting more concerned about their bodies these days.
24	A fifth-grader from Ms. Johnson's class just went to see the school nurse.

Table 15. Filler items for Chapter 3 Experiments 1-3, and the associated target items. Experiment 3 fillers contained the additional text in the first column (if present), while Experiment 1 and 2 fillers used only the original filler text.

Item ID	Target	Extended text	Original filler text
29	scanning		Clara thinks reading the newspaper is a waste of time.
30	binding	Land is not very expensive in the country, but it is near cities.	More and more people own nontraditional homes these days, such as mobile homes.
31	rhane	Our university has an excellent distance learning program.	A lot of students don't attend class. Instead, they watch a video lecture, on the internet.
32	child		I've been to the Carolinas. There are some beautiful mountains there.
33	ghuisery		Civil rights laws must be strengthened. Eventually, the police will stop anyone randomly on the street.
34	phigh		Mark's cabin is surrounded by natural grasslands.
35	wrisc		Betty will be going to the parade this weekend, in honor of the soldiers who have served in the Gulf.
36	ledes	It's pretty hard to find a job in teaching, but there are other options if you want to help people.	I can see a lot of benefits in going into social work as a career.
37	yarreed		The problem with mass media is that it's impossible to educate the public. Anything longer than 30 seconds, they're not interested.
38	joar	The job market is tough.	Irene is planning to join the Peace Corps.
39	shier		Robert gets paid every two weeks. That's 26 paychecks in a year. He pays rent every two weeks also.

Table 15. (Continued)

Item ID	Target	Extended text	Original filler text
40	pe	It used to be that people didn't have children until their twenties.	A lot of parents now are so young. They still don't know what's going on. They don't know what to teach their kids.
41	wocks	I went to that expensive mall on Newbury street, with all the high-end clothing stores.	I got some of those baggy slacks. I think they're called harem pants.
42	galaxy	It's hard to find time to exercise as you get older, I think.	I used to play a lot of soccer in high school, but since I graduated I haven't done much.
43	kares		My husband and I used to live in San Antonio. On the first day of spring last year, there was a huge ice storm.
44	sause	You can't get a dog and then just leave it in the house all day.	Dogs need a lot of attention
45	jail		I worry about my sister. She's a housewife. There's nobody for her to talk to all day except the children.
46	wheeperg		Daniel went to a specialist for a second opinion. That doctor was more optimistic.
47	thwaught		It's great that so many products are made from recycled materials now. Even my parents make sure to recycle.
48	rhodes	The chickens and horses live in the barn all year.	We let the cats go outside as long as it isn't extremely cold. That doesn't happen too often in Texas.
49	muisse		Jacob doesn't like much TV. He does have a couple of favorite shows.
50	phign		They said that we haven't had enough rain this year. That surprises me. It seems like we've had a lot of rain.

Table 15. (Continued)

Item ID	Target	Extended text	Original filler text
51	knix		It's kind of fortunate that the United States was founded in the way that it was, because it's such a mixture of people.
52	momb	The gym I go to is only half a mile from where I live, so it's very convenient	I'm sort of an exercise fanatic. I'm big on swimming.
53	sligh	Action movies are full of terrible acting.	Andrew never watches Steven Seagal movies, ever.
54	briccane		Scientists have been working on speech recognition for a long time.
55	rhuill		The office policy is to earn vacation time. You get a standard number of sick days though.
56	ceast	The company does a lot of checking on its employees.	When I was hired, they made us sign something saying that we could be asked to take a drug test.
57	dersed	New England is a great place to visit.	The best time to come up to Connecticut is in the late fall. The leaves have changed colors. The cold hasn't come in yet.
58	campaign	My family's health history isn't very good, unfortunately.	My grandfather had Alzheimer's disease. My grandmother kept him at home for as long as she could.
59	varf		Sophia lives for spicy food. She cooks everything with peppers.
60	contempt		Michael lives in Oklahoma, about eight miles south of Saint Joe.
61	using	The floor isn't original.	The house was redone before we moved in. It's pretty old. It was built in the fifties.
62	phawner		If you have a kid who wants a pet, it can be a good way to teach them responsibility.
63	gheshed		Nobody here likes violent movies.

Table 15. (Continued)

Item ID	Target	Extended text	Original filler text
64	sa		Lewis spent the whole day at Mount Rushmore waiting to see the fireworks. Unfortunately it got cloudy in the evening.
65	nec		Erica is so happy to have finally found a job. She's going to be an administrator.
66	justice		Those stupid people can't read comic strips. How do you expect them to operate tanks?
67	flame	A funny thing happened when we were out with the kids yesterday.	Two people told us what cute girls we had. We decided it was time to get Mikey his first haircut.
68	ghap		Jennifer would love to visit Egypt. She hopes to take her daughter.
69	whelte		Christine's back has been bothering her. She tries to stay rested. It's hard with the baby.
70	gnip		You shouldn't get your tire plugged if you have a flat. Get it patched instead. It lasts longer.
71	ride		It's time for individuals to start thinking about what they can do to help each other, instead of counting on government.
72	kede	The town is generous all year round, not just at the holidays.	A lot of local stores donate to the community. It's nice of them.
73	renningly	I appreciate the invitation and I'll make a special effort to be on time.	Last time I was at a dinner party, I was late for the actual dinner. I almost missed it.
74	knape	You can make all kinds of things from scratch, and save a lot of money.	It's pretty easy to make fresh pasta. It's very easy if you have a food processor.

Table 15. (Continued)

Item ID	Target	Extended text	Original filler text
75	goniger	Los Angeles has the LA Weekly.	Many cities have alternative newspapers. They have an underground type of feel to them.
76	goldfish	You can't ask job candidates about their health status, or if they go to the doctor.	You can't discriminate against people with certain diseases. The law was passed because of AIDS.